PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES
Volume 18 March 15, 1932 Number 3

ON THE RELATION BETWEEN THE EXPANSION AND THE MEAN DENSITY OF THE UNIVERSE

By A. Einstein and W. de Sitter

Communicated by the Mount Wilson Observatory, January 25, 1932

In a recent note in the *Göttinger Nachrichten*, Dr. O. Heckmann has pointed out that the non-static solutions of the field equations of the general theory of relativity with constant density do not necessarily imply a positive curvature of three-dimensional space, but that this curvature may also be negative or zero.

There is no direct observational evidence for the curvature, the only directly observed data being the mean density and the expansion, which latter proves that the actual universe corresponds to the non-statical case. It is therefore clear that from the direct data of observation we can derive neither the sign nor the value of the curvature, and the question arises whether it is possible to represent the observed facts without introducing a curvature at all.

Historically the term containing the "cosmological constant" λ was introduced into the field equations in order to enable us to account theoretically for the existence of a finite mean density in a static universe. It now appears that in the dynamical case this end can be reached without the introduction of λ.

If we suppose the curvature to be zero, the line-element is

$$ds^2 = -R^2(dx^2 + dy^2 + dz^2) + c^2dt^2, \qquad (1)$$

where R is a function of t only, and c is the velocity of light. If, for the sake of simplicity, we neglect the pressure p,[1] the field equations without λ lead to two differential equations, of which we need only one, which in the case of zero curvature reduces to:

$$\frac{1}{R^2}\left(\frac{dR}{cdt}\right)^2 = \frac{1}{3}\kappa\rho. \qquad (2)$$

The observations give the coefficient of expansion and the mean density:

$$\frac{1}{R}\frac{dR}{cdt} = h = \frac{1}{R_B}; \quad \rho = \frac{2}{\kappa R_A^2}.$$

Therefore we have, from (2), the theoretical relation

$$h^2 = \frac{1}{3}\kappa\rho \tag{3}$$

or

$$\frac{R_A^2}{R_B^2} = \frac{2}{3}. \tag{3'}$$

Taking for the coefficient of expansion

$$h = 500 \text{ km./sec. per } 10^6 \text{ parsecs}, \tag{4}$$

or

$$R_B = 2 \times 10^{27} \text{ cm.},$$

we find

$$R_A = 1.63 \times 10^{27} \text{ cm.},$$

or

$$\rho = 4 \times 10^{-28} \text{ gr. cm.}^{-3}, \tag{5}$$

which happens to coincide exactly with the upper limit for the density adopted by one of us.[2]

The determination of the coefficient of expansion h depends on the measured red-shifts, which do not introduce any appreciable uncertainty, and the distances of the extra-galactic nebulae, which are still very uncertain. The density depends on the assumed masses of these nebulae and on the scale of distance, and involves, moreover, the assumption that all the material mass in the universe is concentrated in the nebulae. It does not seem probable that this latter assumption will introduce any appreciable factor of uncertainty. Admitting it, the ratio h^2/ρ, or R_A^2/R_B^2 as derived from observations, becomes proportional to Δ/M, Δ being the side of a cube containing on the average one nebula, and M the average mass of the nebulae. The values adopted above would correspond to $\Delta = 10^6$ light years, $M = 2.10^{11}\odot$, which is about Dr. Oort's estimate of the mass of our own galactic system. Although, therefore, the density (5) corresponding to the assumption of zero curvature and to the coefficient of expansion (4) may perhaps be on the high side, it certainly is of the correct order of magnitude, and we must conclude that at the present time it is possible to represent the facts without assuming a curvature of three-dimensional space. The curvature is, however, essentially determinable, and an increase in the precision of the data derived from observations will enable us in the future to fix its sign and to determine its value.

[1] It seems certain that the pressure p in the actual universe is negligible as compared with the material density ρ_0. The same reasoning, however, holds good if the pressure is not neglected.

[2] *Bull. Astronom. Inst. Netherlands, Haarlem*, **6**, 142 (1931).

AZO DERIVATIVES OF THE PYRIMIDINES*

BY MARSTON T. BOGERT AND DAVID DAVIDSON

DEPARTMENT OF CHEMISTRY, COLUMBIA UNIVERSITY

Communicated February 15, 1932

Introduction.—This paper contains a preliminary survey of the azo derivatives of the pyrimidines with particular reference to derivatives of uracil. These derivatives fall into five natural types according to the nature of the components entering into their formation as follows:

Those formed by

(1) The coupling of aromatic diazonium salts with pyrimidine phenols.

(2) The coupling of diazopyrimidines with aromatic phenols.

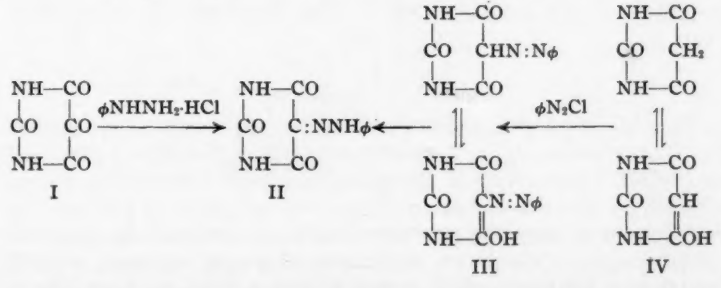
(3) The coupling of diazopyrimidines with pyrimidine phenols.

(4) The condensation of aromatic hydrazines with 5-keto-6-hydroxy-5,6-dihydropyrimidines.

(5) The condensation of hydrazinopyrimidines with 5-keto-6-hydroxy-5,6-dihydropyrimidines.

Examples of each of these types of azo derivatives had already been studied by us when the paper of Johnson, Baudisch and Hoffmann[1] on diazouracil appeared. Since these authors have indicated their intention of continuing the study of derivatives of this diazopyrimidine, we are publishing the results of our preliminary work at this time.

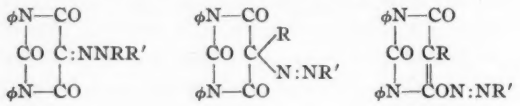
Azopyrimidines of Type 1.—Pelizzari[2] was the first investigator who attempted to condense alloxan (I) with phenylhydrazine. He failed owing to the occurrence of an oxidation-reduction reaction in which the alloxan was reduced to alloxantine and the phenylhydrazine oxidized to benzene and nitrogen. By using phenylhydrazine hydrochloride rather than the base, Kühling[3] succeeded in obtaining alloxan-5-phenylhydrazone



(II). Several years later, Kühling coupled benzene diazonium chloride with barbituric acid (IV) and found the product to be identical with alloxan-5-phenylhydrazone. II and III are, therefore, desmotropes of

the quinonehydrazone-azophenol type. The isolated material appears to be represented by formula II, since the 1,3-dialkyl derivatives are not soluble in alkali. Kühling also prepared the ortho- and para-nitro derivatives of II.

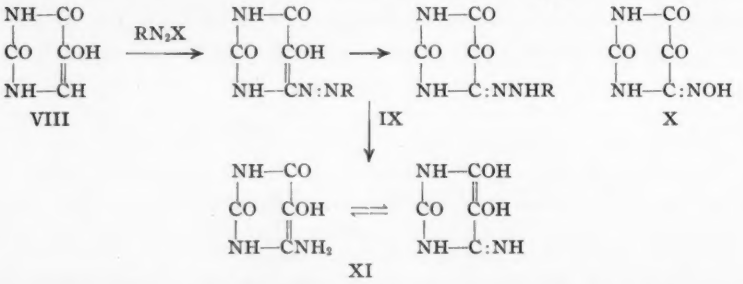
Whitely[4] coupled benzene diazonium chloride with 1,3-diphenyl barbituric acid and found the product to be identical with that obtained by condensing phenylhydrazine with 1,3-diphenyl-5,5-dibromobarbituric acid (1,3-diphenylalloxan-5-phenylhydrazone). This author confirmed the idea that formula II represents the structure of the alloxan-5-hydrazones by preparing compounds typified by formula V, which proved to be quite similar to the former in their properties.



Whitely also claims to have coupled aromatic diazonium salts with 1,3-diphenyl-5-alkylbarbituric acids, obtaining products typified by formula VI. These products, however, exhibit the characteristic instability of diazonium salts, and should perhaps better be considered as such. Formula VII represents these products as diazonium barbiturates.

Nightingale and Schaefer[5] have recently described the preparation of alloxan-5-α-naphthylhydrazone by the condensation of α-naphthylhydrazine with 5,5-dichloro- or dibromobarbituric acids.

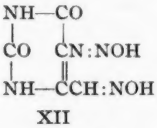
We have found that isobarbituric acid (VIII) readily couples in weakly acid solution with aromatic diazonium salts to give highly colored products



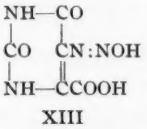
which appear to have the structure typified by formula IX (alloxan-6-arylhydrazones).[6] These new derivatives of alloxan represent unstable derivatives of this cycle, which cannot be obtained directly from alloxan. Like the corresponding oxime (isovioluric acid or alloxan-6-oxime) (X) described by Davidson and Bogert,[7] they are capable of condensing with carbonyl reagents because of the presence of the reactive 5-carbonyl group

of alloxan. They are rearranged by acids to the stable alloxan-5-arylhydrazones, while they are slowly decolorized in alkaline solution (probably owing to ring scission). Reduction yields a mixture of isouramil (XI)[8] and an aromatic amine. While the alloxan-5-arylhydrazones containing only one phenyl group are generally yellow, the corresponding 6-arylhydrazones are deep red.

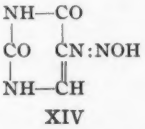
Azopyrimidines of Type 2.—Behrend and Ernert[9] showed that 5-amino derivatives of uracil (5-amino-6-methyluracil and 5-aminouracil-6-carboxylic acid) could be diazotized. Angeli[10] diazotized 5-aminouracil itself and thus obtained 5-diazouracil which was also secured by Behrend and Ernert in the decarboxylation of diazotized 5-aminouracil-6-carboxylic acid. From the analyses of these diazopyrimidines Behrend and Ernert provisionally proposed the structures XII, XIII, and XIV, respectively.



XII

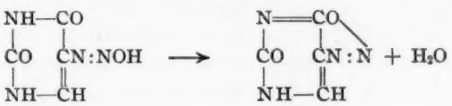


XIII



XIV

These authors found that these compounds, with the exception of XIII, which could not be dried without losing CO_2, could be obtained in anhydride forms. They considered the loss of water to occur between the diazonium group and one of the NH groups, but since they could not determine which of the two NH groups was involved, they preferred to represent these diazopyrimidines as given, rather than as diazoanhydride hydrates. Johnson, Baudisch, and Hoffmann[11] also obtained the anhydride of XIV and considered it to be formed by the loss of water between the diazonium group and the enolized NH—CO group in the 3,4 position:



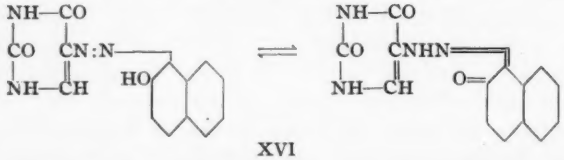
XIV XV

These authors convey the impression that this anhydride, $C_4H_2O_2N_4$, is the product which they obtained directly from the diazotization. Inspection of their published analyses shows, however, that $C_4H_2O_2N_4$ was only obtained by drying at 100°, and that the loss of weight on drying at 100° corresponds to that required by the reaction given above. (Calcd. for the loss of H_2O from $C_4H_4O_3N_4$: 11.5%. Found (mean): 11.5%.)

It appears, therefore, that these authors had in hand the same product which Behrend and Ernert had. The difference in color which they felt distinguished their product from Behrend and Ernert's is not valid, since Behrend and Ernert described diazouracil as yellowish white, yellow, or

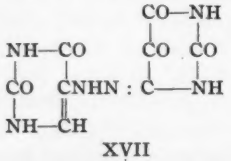
red depending on the vigorousness of the method by which they produced it. It thus remains an open question whether diazouracil ($C_4H_4O_3N_4$) is to be represented as in XIV or as a hydrate of XV. Since, as Behrend pointed out, diazonium compounds in the aromatic series generally appear as salts of the acids in which they are produced, except when an acid group is present in the amine molecule (as in sulfanilic acid) when they form internal salts or diazoanhydrides, it is reasonable to suppose that 5-aminouracil, containing as it does the acidic NH—CO (or N:COH) group,[12] should also give rise to a diazoanhydride. We may, therefore, assume $C_4H_4O_3N_4$ to be a hydrate of XV.

Behrend reported that diazotized 5-amino-6-methyluracil (5-diazo-6-isonitrosomethyluracil) (XII) did not couple with phenols and amines, while no mention was made of the coupling power of XIII or XIV. More than two years ago we first made the qualitative observation that diazouracil couples with aromatic phenols in alkaline solution, and we have now isolated several of these products (from α-naphthol, β-naphthol, and R-salt). Johnson, Baudisch and Hoffmann reported the coupling with α-naphthol, although no analysis was given. These azo derivatives may be represented thus (with β-naphthol, for example):



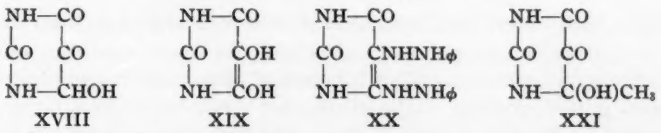
XVI

Azopyrimidines of Type 3.—Thus far, we have considered the coupling of aromatic diazonium salts with pyrimidine phenols and the coupling of diazopyrimidines with aromatic phenols. With the experience thus gained it is not surprising to find that diazopyrimidines may be coupled with pyrimidine phenols. Thus, diazouracil couples with isobarbituric acid in alkaline solution or in glacial acetic acid to give a brown product, the sodium salt of which is chocolate-brown with a purple reflex and gives a brilliant red solution in water. This derivative is to be formulated thus:



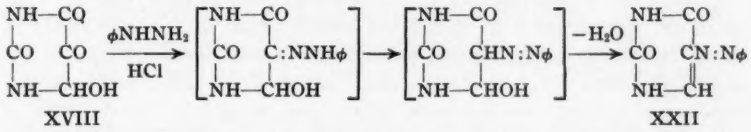
XVII

Azopyrimidines of Type 4.—Of the known 5-keto-6-hydroxy-5,6-dihydrouracils, two have been of particular interest. These are isodialuric acid (XVIII)[13] and 6-methylisodialuric acid (XXI).[14]



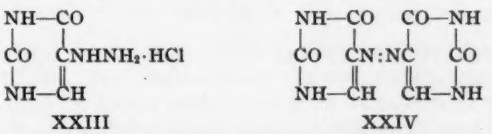
Concerning the action of phenylhydrazine on these derivatives, Behrend and his students found that 6-methylisodialuric acid gave a phenylhydrazine salt when treated in aqueous solution with phenylhydrazine. Levene[15] found that isodialuric acid was converted into 5,6-di-phenylhydrazinouracil (XX) by the action of a hot acetic acid solution of phenylhydrazine. This result becomes understandable when we recall that Behrend, Koch, and von Vogel[16] noticed that in hot acetic acid even bases like guanidine and thiourea isomerized isodialuric acid into dialuric acid (XIX).

By employing phenylhydrazine hydrochloride, we have been able to avoid this isomerization and to obtain a direct condensation between isodialuric acid and phenylhydrazine. The reaction proceeds further, however, with the elimination of a second molecule of water, the ultimate product being the orange-brown *uracil-5-azobenzene* (XXII). Below is given a hypothesis for the mechanism of this reaction:[17]



Uracil-5-azobenzene exhibits the characteristic behavior of aromatic azo compounds toward such reagents as sodium sulfite, hydrochloric acid, potassium iodide, and reducing agents.

Azopyrimidines of Type 5.—By extending the reaction just discussed to 5-hydrazinouracil (XXIII) it was anticipated that azouracil-5,5′ (XXIV) would be obtained. In agreement with this expectation, the product of the reaction was orange-yellow in color and behaved typically toward the reagents used in characterizing aromatic azo compounds. At the present writing, however, we have not overcome certain difficulties in the analysis of the product, and hence cannot report its constitution with certainty.



Experimental.—Coupling of aromatic diazonium salts with pyrimidine phenols (type 1). A solution of 0.01 mol of aromatic amine in 0.03 mol

of dilute hydrochloric acid was treated with 0.01 mol of sodium nitrite. 0.03 mol of sodium acetate was then added and finally a solution of 0.01 mol of, say, isobarbituric acid in 0.02 mol of dilute sodium hydroxide was dropped in with stirring. After stirring for a half hour, the product was separated by filtering or centrifuging. The yields throughout this paper ranged from 70–90%.

Alloxan-6-phenylhydrazone is insoluble in water or alcohol. It is soluble in glacial acetic acid from which it crystallizes in red needles, m. 260° (dec.). It forms a red solution in alkali which soon becomes decolorized.

Anal. Calcd. for $C_{10}H_8O_2N_4$: C, 51.7; H, 3.5%. Found: C, 51.7; H, 3.4%.

Alloxan-6-phenylhydrazone-p-sodium sulfonate, chocolate-brown powder giving a red solution in water. It dyes wool a Burgundy red. When heated at 100° for several hours with two equivalents of dilute hydrochloric acid, it is rearranged to the product given below.

Anal. Calcd. for $C_{10}H_7O_6N_4SNa \cdot 1^1/_2H_2O$: C, 33.2; H, 2.8%. Found: C, 33.3, 33.1; H, 2.9, 2.8%.

Alloxan-5-phenylhydrazone-p-sodium sulfonate, bright yellow needles, dyes wool yellow.

Anal. Calcd. for $C_{10}H_7O_6N_4SNa$: C, 35.9; H, 2.1%. Found: C, 35.5, H, 2.4%.

Coupling of diazouracil with aromatic phenols and amines (type 2). Diazouracil was prepared by dissolving 5-aminouracil in 1.1 equivalents of dilute hydrochloric acid, adding ice and then a solution of 1.1 equivalents of sodium nitrite. The product separated at once as a buff-colored powder, which was filtered, washed with water, and air dried. In coupling, a suspension of 0.01 mol of diazouracil in water was added to an iced solution of 0.01 mol of the phenol in 0.02 mol of dilute sodium hydroxide and the solution stirred for several hours. Generally, an insoluble sodium salt separated, which was treated with acetic acid to obtain the free azo derivative. In an alternative method of coupling, 0.01 mol of diazouracil was suspended in a solution or suspension of 0.01 mol of the phenol and 0.02 mol of anhydrous sodium acetate in 20 cc. of glacial acetic acid and the mixture shaken for several hours.

Uracil-5-azo-α-naphthol, dark reddish-brown powder, insoluble in water and benzene, soluble in alcohol, glacial acetic acid, and alkalies, m. above 300°. For analysis it was crystallized from alcohol.

Anal. Calcd. for $C_{14}H_{10}O_2N_4$: C, 59.6; H, 3.5%. Found: C, 59.7; H, 3.8%.

Uracil-5-azo-β-naphthol, scarlet powder, insoluble in water or benzene, soluble in alcohol, glacial acetic acid, and alkalies, m. 285° (dec.).

Anal. Calcd. for $C_{14}H_{10}O_2N_4$: C, 59.6; H, 3,5%. Found: C, 59.6; H, 3.8%.

Uracil-5-azo-β-naphthol-3,6-disodium sulfonate (*R-salt*), red needles with a metallic lustre, dissolves in water to give a red solution, dyes wool a deep orange-red.

Anal. Calcd. for $C_{14}H_8O_9N_4S_2Na_2 \cdot 3H_2O$: C, 31.1; H, 2.6%. Found: C, 31.1; H, 2.7%.

Coupling of diazouracil with pyrimidine phenols (type 3). This reaction occurred in alkaline solution or in glacial acetic acid containing sodium acetate as described under type 2.

Uracil-5-azoisobarbituric acid (alloxan-6-(uracil-5-)hydrazone), chocolate-brown powder, insoluble in water and organic solvents, soluble in alkalies to form a deep red solution which soon fades. It does not melt below 300°.

Anal. Calcd. for $C_8H_6O_6N_6$: C, 36.1; H, 2.3%. Found: C, 35.8; H, 3.2%.

Uracil-5-azobarbituric acid (alloxan-5-(uracil-5-)hydrazone), orange-yellow powder, insoluble in water and organic solvents, soluble in alkalies, does not melt below 300°.

Condensation of hydrazine salts with isodialuric acid (types 4 and 5). To a solution of 0.01 mol of isodialuric acid in 15 cc. of water was added a solution of 0.01 mol of the hydrazine salt containing a drop of hydrochloric acid. The mixture turned orange at once and a precipitate began to form. The reaction was complete within 24 hours. The product was filtered, washed with water, and dried at 100°.

Uracil-5-azobenzene, orange-brown powder, slightly soluble in cold water, alcohol, and glacial acetic acid, but readily soluble hot; also soluble in concentrated sulfuric acid, concentrated hydrochloric acid, alkalies, and sodium sulfite; liberates iodine from acidified potassium iodide; on reduction with aluminum and sodium hydroxide it yields aniline and 5-aminouracil.

Anal. Calcd. for $C_{10}H_8O_2N_4$: C, 55.5; H, 3.7; N, 26.0%. Found: C, 55.4; H, 3.6; N, 26.1%.

Uracil-5-azobenzene-p-sodium sulfonate, light brown needles, by salting out the reaction product of phenylhydrazine-*p*-sulfonic acid on isodialuric acid, soluble in water, dyes wool orange.

* Contribution from the Chemical Laboratories of Columbia University, with the coöperation of the Research Department of the American Manufacturers of Toilet Articles.

[1] Johnson, Baudisch, and Hoffmann, *Ber. Chem. Ges.*, **64**, 2629 (1931).

[2] Pelizzari, *Gazz. chim. ital.*, **17**, 254 (1887).

[3] Kühling, *Ber. chem. ges.*, **24**, 4140 (1891); **31**, 1972 (1898).

[4] Whitely, *J. Chem. Soc.*, **91**, 1330 (1907).

[5] Nightingale and Schaefer, *J. Am. Chem. Soc.*, **54**, 236 (1932).

[6] Johnson and Jones, *Am. Chem. J.*, **40**, 538 (1908), noted the production of a red color from the action of diazobenzene-*p*-sulfonic acid on isobarbituric acid in the presence of sodium bicarbonate. They neither isolated the product nor proposed a structure for it, but related it to the colors obtained by Johnson and Clapp, *J. Biol. Chem.*, **5**, 163 (1908), by the action of diazobenzene-*p*-sulfonic acid on all uracil derivatives in which the 1-position was unsubstituted. [In Johnson's papers this position is numbered "3." The numbering in the present paper conforms with the system adopted by *Chemical Abstracts*.]

[7] Davidson and Bogert, forthcoming publication.

[8] Davidson and Bogert, forthcoming publication.

[9] Behrend, *Ann. Chem.*, **245,** 213 (1888); Behrend and Ernert, *Ibid.*, **258,** 347 (1890).

[10] Angeli, *Atti. accad. Lincei*, [5] **3,** II, 72 (1894).

[11] Johnson, Baudisch, and Hoffmann, loc. cit.

[12] Levene, Bass, and Simms, *J. Biol. Chem.*, **70,** 229 (1926).

[13] Behrend and Roosen, *Ann. Chem.*, **251,** 235 (1889).

[14] Behrend and Osten, *Ann. Chem.*, **343,** 133 (1905); Behrend and Beer, *Ibid.*, **362,** 115 (1908).

[15] Levene, *J. Biol. Chem.*, **63,** 653 (1925).

[16] Behrend and Koch, *Ann. Chem.*, **315,** 246 (1901); Behrend and von Vogel, *Ibid.*, 259 (1901).

[17] This reaction suggests a mechanism for the formation of 5-phenylhydrazinouracil from the action of phenylhydrazine on 5,5-dibromo-6-hydroxy-5,6-dihydrouracil which was observed by Levene: namely, the formation of uracil-5-azobenzene (much as given above) followed by its reduction to the hydrazo compound by the excess of phenylhydrazine present.

SOME EFFECTS OF HIGH TEMPERATURE ON POLYPLOIDY AND OTHER VARIATIONS IN MAIZE[1]

BY L. F. RANDOLPH

UNITED STATES DEPARTMENT OF AGRICULTURE AND CORNELL UNIVERSITY

Communicated February 1, 1932

A renewed interest in the study of variations followed the recent discovery that profound changes in the heredity of organisms may be induced by x-radiation. Intensive studies in this field have demonstrated effects, some of which appear to be individual gene mutations modifying very slightly or not at all the organization of the chromosome as a morphological unit; others are known definitely to involve groups of genes and are exhibited not infrequently as visible structural changes in the architecture of the chromosome. These latter chromosomal variations have led to a better understanding of the probable mode of origin of morphologically unlike chromosomes in related species.

Variations of still another kind occur in nature and involve the duplication or loss of entire chromosome sets. From such variations have originated the polyploid series of multiple chromosome numbers, which are to be found very generally among related species distributed throughout the plant kingdom. The study of this class of variations has progressed slowly because of their infrequent occurrence under natural conditions. A technique for the production of new forms with increased or reduced numbers of chromosome sets would make possible the development of new

and improved varieties through the removal of incompatibilities of the type due to differences in chromosome number, and would provide much favorable material for other cytogenetic investigations of species relationships.

Numerous attempts have been made to induce polyploidy, and in certain cases a measure of success has been attained. Noteworthy are the classical experiments of the Marchals with mosses and of Winkler with Solanum. These and other more recent investigations have demonstrated the possibility of inducing polyploidy in somatic tissues by wounding, grafting, etc.[2] Such methods are, however, restricted in their applicability chiefly to plants which may be propagated asexually. The search for a method of wider applicability has continued.

The effectiveness of abnormally high temperatures in producing tetraploid cells and sectors in somatic tissues, chiefly of root-tips, has been noted by workers interested primarily in determining the reaction of the chromosomes to various physical and chemical agents (Nemec,[3] Sakamura,[4] Lundegårdh[5]). Sakamura observed an increased frequency of tetraploid cells in root-tips of Pisum which had been treated with hot water at 40 degrees Centigrade. Similar results with somewhat higher temperatures were reported by Koshuchow[6] working with the root-tips of Cucumis and Zea. Irregular chromosome division and fragmentation resulting from high temperature treatments were observed by Lundegårdh[5] in Vicia roots.

The relation of high temperature to gene mutation has been studied in Drosophila by a number of investigators. Muller[7] compared the frequencies of lethal mutations in cultures of flies held at 19 degrees Centigrade and at 27 degrees Centigrade, and noted a slightly higher rate in the warmer series. The difference was considered to be significant although the cultures used in the experiments showed a high variability in mutation rate. Goldschmidt,[8] Jollos[9] and Rokizky[10] worked with much higher temperatures (37 degrees Centigrade) and reported markedly increased frequencies of induced mutations. Many of these were identical with previously known mutations and occurred repeatedly in their experiments.

In a series of high temperature experiments undertaken several years ago for the primary purpose of developing a technique for producing tetraploids in maize, preliminary tests were made to determine the effects of high temperatures on cell division in seedling root-tips. Seedlings with well developed seminal roots were immersed in hot water at from 40 to 45 degrees Centigrade for different periods ranging from 15 minutes to two hours, and were maintained thereafter under conditions favorable for growth. Root-tips were fixed at frequent intervals during the 24-hour period following treatment. Microscopical examination of the root meristems showed an almost complete cessation of growth immediately following treatment, as indicated by the absence of division figures. Division activity was

TABLE 1

KERNEL TYPES AND SEEDLING CHROMOSOME NUMBERS IN THE HIGH TEMPERATURE EXPERIMENTS

EXPT.	SCHEDULED TREATMENT TEMPERATURE	KERNEL TYPES: NORMALLY DEVELOPED KERNELS SELF-COLORED	NORMALLY DEVELOPED KERNELS CHIMERAS*	DEFECTIVE KERNELS: PARTIALLY DEVELOPED ENDOSPERMS	DEFECTIVE KERNELS: GERMLESS	DEFECTIVE KERNELS: SCARRED ENDOSPERMS	TOTAL	SEEDLING CHROMOSOME NUMBERS: 2N	4N	4N CHIMERAS	8N	TOTAL
A	45	722	21	73	87	89	903	567	20	7	1	595
B	43	284	..	14	61	14	359	215	5		1	221
C	43	169	9	13	17	13	208	142	3	1	1	147
D	40	471	59	32	46	32	618	374	1	2		377
E	38	359	7	8	16	8	390	312		2		314
Total		2005	96	140	227	156	2478	1610	29	12	3	1654

* Only chimeras in which the area lacking the dominant character or group of characters involved more than one-sixteenth of the surface area of the endosperm were recorded.

slowly resumed, ordinarily after about one hour, with a complete recovery of the normal rate within two or three hours after treatment. Tetraploid sectors were found to be much more abundant in treated roots fixed 24 hours after treatment than in the untreated roots in the same experiments. In an experiment which may be cited as representative of a series of similar experiments, 104 roots treated at 45 degrees Centigrade for one hour were examined for tetraploidy and 15 were found to contain one or more sectors of $4n$ tissues. Among 84 untreated roots in the same experiment only one contained a $4n$ sector. Successive treatments applied to the roots at four hour intervals materially increased the number of $4n$ sectors in a given number of roots.

Having thus obtained an effective technique for inducing tetraploidy in root meristems, similar treatments were applied to the zygote and proembryo in an attempt to produce entire individuals with multiple chromosome numbers. This was accomplished in Experiments A and B (Table 1) by applying heat locally to the ear-shoots. The ear-shoot region of the plant was enclosed within a cylinder of wire mesh. The cylinder was plugged at the ends with cotton and surrounded by an electrical heating pad. The temperature of the ear-shoot was determined approximately by a thermometer placed against the side of the ear within the husks, the temperature being regulated by varying the amount of current passing through the heating element of the pad. A thermometer inserted in the

air space inside the cylinder recorded the air temperature about the ear. In Experiments C, D and E, the entire plants were placed for treatment in a room maintained at the desired temperature.

The method of applying the heat treatments consisted in maintaining an air temperature of 47–48 degrees Centigrade until the ear-shoot reached the scheduled treatment temperature (column 2, Table 1). After this temperature was reached it was maintained with only minor fluctuations for approximately an hour by adjusting the air temperature, following which it was allowed to revert to the normal field or greenhouse temperature. Since temperatures higher than 48 degrees often were injurious, refinements of the present method or entirely new methods of maintaining high temperatures within the tissues of the plant would have to be devised in order to test the effects of temperatures appreciably higher than 45 degrees Centigrade.

The ear-shoot treatments were applied in Experiments A, C and E during the 48-hour period beginning 27–30 hours after pollination, this being the interval during which the division of the zygote and the early divisions of the proembryo were observed to take place under the conditions of temperature which were maintained during the experiments. In Experiments B and D the treatments were begun earlier, at 22–24 hours after pollination, to include the period during which syngamy occurred. The treatments were repeated at regular four-hour intervals throughout the entire 48-hour period in the initial experiments. In later experiments treatments were confined to the period of rising temperature in the morning and early afternoon when cell-division activity was at a maximum, no treatments being given during the night when division activity was at a minimum. Three treatments usually were applied during each of the two or three days following that on which the pollinations were made. Thirty plants were treated in Experiment A, eight in Experiment C and 12 in each of the other three experiments.

The stocks used in the experiments were selected to provide material for a study of the mode of origin and other characteristics of tetraploids, and also to afford a means of identifying genetically other types of chromosomal alterations and to make possible the recognition of disturbances in the normal processes associated with syngamy. For Experiment A, and F_1 hybrid between two self-fertilized strains having colorless aleurone, yellow endosperm, and weak anthocyanin color in the seedling stage was pollinated by a homozygous sun-red, liguleless type with purple aleurone and colored scutellum $\left(\frac{A\ C\ r^{rw}\ sc_x\ b-Lg\ pl-Y}{A\ C\ R^r\ Sc_x\ B-lg\ pl-y}\right)$. For Experiment B, a self-fertilized strain of a commercial variety was used and the plants were self-pollinated. For Experiment C, a self-fertilized strain with colorless aleurone, white endosperm, and lacking anthocyanin color in the

seedling stage was crossed with the same pollen parent that was used in Experiment A $\left(\frac{A\ C\ r^g\ sc_x\ b-Lg}{A\ C\ R^r\ Sc_x\ B-lg}\right)$. For Experiments D and E, a liguleless type lacking anthocyanin color in the seedling stage, heterozygous for the *C*-aleurone factor, and for shrunken, waxy, sugary, yellow endosperm was used as the pistillate parent. Supernumerary *B*-type chromosomes[11] had been introduced previously into this stock, and for these experiments plants with one or more, preferably an odd number, of *B*-type chromosomes were selected. A dilute sun-red, glossy seedling type with sugary, yellow endosperm was used as the pollen parent $\left(\frac{a\ \ b-lg\ \ R^g\ pl-Yy\ \ Gl_3-Susu\ Cc-Shsh-Wxwx}{A\ B-Lg\ r^r\ \ pl-YY\ gl_3-susu\ \ CC-ShSh-WxWx}\right)$. The writer is indebted to Professor R. G. Wiggans of Cornell University for the self-fertilized strains of commercial varieties, and to Dr. G. F. Sprague of the United States Department of Agriculture for the stock homozygous for scutellum color.

Effects of the heat treatments were noted as follows:

1. Doubling of entire chromosome sets.
2. Chromosomal deficiencies and translocations.
3. Direct morphological effects resulting in defective and scarred endosperms, germless grains, dwarfed and otherwise defective seedlings and mature plants.
4. Sterilization of ovules and pollen at the higher temperatures.
5. Deviations from the normal fertilization process, such as failure of syngamy, atypical syngamic unions and parthenogenesis.

The polyploids were detected by a microscopical examination of seedling radicles.[12] These determinations were verified in all cases by examining additional root-tip material, and in Experiment A were further supplemented by sporocyte and pollen examinations, tests of the behavior of the pollen in crosses with diploids, and by noting the segregation of characters in the selfed progeny. In Experiment A, 17 of the 20 plants having $4n$ roots were found by these later tests to be wholly tetraploid, the others being chimeras with $4n$ roots and $2n$ stem portions. The $4n$ chimeras listed in Table 1 were seedlings with radicles containing both $2n$ and $4n$ tissue, one-eighth or more being tetraploid.

The higher temperatures were more effective than the lower in producing tetraploids and aberrant kernel types, as shown in table 1. An apparent exception is Experiment D, in which a high frequency of endosperm chimeras, germless and otherwise defective kernels was associated with a relatively low treatment temperature and a low frequency of chromosome doubling. In this experiment, however, the ear-shoots received in addition to the heat treatments approximately 500 r-units of x-radiation 23 hours

after pollination. The high frequency of aberrant kernel types in this experiment probably was caused by the x-rays, since it has been shown that they produce alterations of this kind.[13] Attempts to induce chromosome doubling with x-rays have failed repeatedly in experiments with maize.

All of the endosperm chimeras in which the dominant character or characters were absent from one-sixteenth or more of the surface area of the endosperm and present in the remaining portion are included in table 1. Most of these probably were due to individual chromosome deficiencies, as demonstrated originally by Emerson[14] for untreated material and similarly interpreted by Stadler[15] for x-rayed material. Three hybrid grains with colorless aleurone and yellow endosperm and five non-liguleless seedlings lacking anthocyanine color, which occurred in Experiments A and C, also may be classed as chromosome deficiencies induced at a sufficiently early stage to produce entire endosperms and embryos with the deficiency. These seedling recessives were deficient in growth and partially sterile. The frequency of chromosome translocations, associated with partial sterility and altered synaptic behavior of the chromosomes in meiosis, was materially increased by the heat treatments. Among 139 individuals with normal vigor there were 11 semi-sterile plants and a smaller number which exhibited other larger percentages of sterility.

A study of the effects upon the fertilization process of heat treatments applied separately and in combination with x-radiation at the time of syngamy was undertaken in Experiments A and D, and the following variations were noted:

1. A new type of endosperm chimera, not previously reported in maize, originated from the cross *aa RR Cc−Shsh−Wxwx Yy Susu* × *AA rr CC−ShSh−WxWx YY susu*. Approximately one-half of the endosperm of this exceptional kernel was maternal in character, exhibiting the recessive colorless aleurone and waxy, non-yellow endosperm characters, and the dominant, non-sugary endosperm character contributed by the female parent. The remainder of the endosperm was hybrid in nature as indicated by colored aleurone, the complementary action of the three genes A, C and R being requisite to the production of colored aleurone. This portion of the grain also had non-waxy, yellow, non-sugary endosperm. This chimera cannot be interpreted as a single chromosome loss, but may be interpreted readily by assuming that one polar nucleus developed independently to form the maternal portion and the second polar nucleus fused with the second male nucleus to form the hybrid portion. Thus the colorless portion would be haploid and the colored portion diploid. The aleurone cells in the transition region between the colored and colorless portions were examined microscopically and those of the latter were found to be smaller than those of the former, thus adding further confirmation to this interpretation.

2. A diploid glossy seedling was produced in Experiment D from a colored starchy grain. If the glossy character alone were considered, this plant might be interpreted merely as a *Gl* deficiency, but the fact that it differed markedly from seedlings of this type in having normal vigor and in resembling in general appearance the male parental strain rather than the normal hybrid seedlings, suggested that it was a paternal diploid. If evidence from the selfed progeny of this plant, which has not yet reached maturity, becomes available, its mode of origin can be more definitely established.

3. A maternal diploid seedling also was produced in Experiment D. The parthenogenetic development of a haploid egg followed by an early doubling of its chromosomes is indicated as the most probable mode of origin of this exceptional seedling.

4. A maternal haploid appeared in Experiment B. Although this haploid and the three exceptional individuals already noted occurred among relatively few individuals in an experiment designed primarily to test the effects of heat treatments on the fertilization process, more data are required to establish a definite causal relation.[15] The evidence from the preceding three exceptional cases is somewhat better in that individuals of these types have not been reported previously in maize.

The effectiveness of heat treatments in producing gene mutations was studied in Experiment A. Progenies consisting of 25 or more individuals were grown from the self-pollinated ears of 113 plants treated in the initial stages of embryogeny. These plants were from six ears selected from among 12 fertile ears in this experiment as showing the most pronounced effects of the heat treatments. These progenies consisting of 3975 seedlings were critically examined for new characters. Induced recessive mutations, as well as characters originally present, should appear in these families. The chance that one or more individuals recessive for a character segregating as a simple mendelian recessive will appear in a random F_2 population of 25 individuals is 200 to 1. Two seedlings showing irregular chlorophyll striping similar to that observed in cases of maternal inheritance in maize were found in one family. In 15 of 33 families, all of which were from the same original cross, one or more glossy seedlings were found. The recurrence of this glossy mutant in approximately one-half of the F_2 progenies of one original cross strongly suggests that one parent was heterozygous for the character. The low frequency of occurrence of segregating mutants in these F_2 families was probably due to the fact that the stocks used in the experiments were from self-fertilized strains which were more nearly homozygous than the stocks ordinarily used in genetic studies in maize. Certainly there was no evidence from these experiments that high temperatures are effective in causing gene mutation, but tests on a larger scale might establish a causal relation.

The genetic changes induced by the high temperatures were expressed in modifications of, first, the processes associated with nuclear and cell division, and, second, the architecture of individual chromosomes. The former make possible among other things the experimental production of new polyploid strains, the duplication of the polyploid condition in existing species, the production of hybrids between species incompatible because of differences in chromosome number by equalizing the parental chromosome numbers before crossing, and the production of fertile tetraploid hybrids between species whose normal diploid hybrids are sterile. The latter provides a new technique for a study of the physical properties of individual chromosomes, and should aid materially in the search for the underlying causes of induced chromosomal variations. Since organisms in their natural environment are not infrequently subjected to temperatures as high as or higher than those employed in these experiments, it seems not unlikely that high temperatures have played an important part in the production of variations in nature.

[1] These experiments were aided by a grant from the NATIONAL RESEARCH COUNCIL, Committee on Effects of Radiation upon Living Organism.

[2] A résumé of this work was published by Ufer, Max, *Der Zuchter*, **1**, 225–230(1929).

[3] Němec, D., *Das Problem der Befruchtungs-vorgänge*, Berlin (1910).

[4] Sakamura, T., *J. Coll. Imp. Univ. Tokyo*, **39**, pp. 221 (1920).

[5] Lundegårdh, H., *Svensk. Bot. Tidskr.*, **8**, 161–180 (1914).

[6] Koshuchow, Z. A., *Zeitschr. für Erforschung der Nutzpflanzen*, **10**, 140–148 (1928).

[7] Muller, H. J., *Genetics*, **13**, 279–357 (1928).

[8] Goldschmidt, R., *Biol. Zentralbl.*, **49**, 437–448 (1929).

[9] Jollos, Victor, *Ibid.*, **50**, 541–554 (1930).

[10] Rokizky, P. Th., *Ibid.*, **50**, 554–566 (1930).

[11] Randolph, L. F., *Anat. Record*, **41**, 102 (1928).

[12] Subsequently it was discovered that the tetraploids and other members of the polyploid series could be identified by noting the size and distribution of the stomata in the seedling stage. The stomata of haploids were found to be definitely smaller and closer together than were those of diploids, while in the triploids, tetraploids and octoploids they were correspondingly larger and more widely separated. These differences may be detected at relatively low magnifications, as with a 20X hand lens.

[13] Stadler, L. J., these PROCEEDINGS, **16**, 714–720 (1928), and unpublished data.

[14] Emerson, R. A., *Amer. J. Bot.*, **8**, 411–424 (1921).

[15] Stadler, L. J., loc. cit.

[16] Haploidy in maize was first reported by L. J. Stadler, and the writer, in papers presented before Section O, Amer. Assoc. Adv. Sci., Des Moines, Iowa, 1929. (Unpublished.)

NOTE ON THE QUANTUM DYNAMICAL CORRECTION OF THE EQUATION OF STATE

By Henry Margenau

Sloane Physics Laboratory, Yale University

Communicated February 4, 1932

We desire to correct a numerical error which occurred in a recent paper bearing this title.[1] In the definition of $[\rho^2]$, following eq. (11), m should be replaced by $m/2$, the reduced mass in the relative motion of two molecules. This requires a change in the argument of φ from $-\frac{\epsilon}{kT}$ to $-\frac{2\epsilon}{kT}$, if φ is defined as in the paper. The final formula (15) remains as stated if the numerical coefficients appearing in $\varphi(x)$ are changed correspondingly. Thus

$$\varphi(x) = 0.752x^{3/2} - 0.451\,x^{5/2} + 0.161\,x^{7/2} - 0.042\,x^{9/2} + \ldots$$

if $x > 0$.

Meanwhile there appeared a paper on the same subject by Kirkwood.[2] He corrects a formula previously proposed by Kirkwood and Keyes[3] and arrives at a result which differs somewhat from ours. To compare the two it should be observed that our $\varphi(x)$ is the same as Kirkwood's $\left\{G(x^{1/2}) - \frac{2}{\sqrt{\pi}}\,x^{1/2}\,e^{-x}\right\}$ and is perhaps better stated in this closed form, although the series is found very convenient for numerical computations. We do not, however, understand his objection to our formula. He remarks that it gives a value smaller than the classical B at higher temperatures. This is not the case. The term $\varphi(x)$ renders smaller only the negative contribution of the integral to B, as an inspection will show. For high temperatures our formula goes over into the classical expression for B.

We cannot refrain from commenting on Kirkwood's numerical calculation of the second virial coefficient for He. As he points out, his expression (23) for ϵ is obtained by adding to the resonance energy that portion of the Van der Waals energy which corresponds to dipole interaction only, leaving out of account the quadrupole term which has been shown to be appreciable.[4] Its insertion would have destroyed the agreement with experiment. Thus the agreement with empirical data appears to have been produced somewhat *ad hoc*, and the theoretical values of B do not seem convincing.

[1] These Proceedings, **18**, 56 (1932).

[2] J. G. Kirkwood, *Phys. Zeits.*, **33**, 39 (1932).

[3] J. G. Kirkwood and F. G. Keyes, *Phys. Rev.*, **38**, 516 (1931).

[4] H. Margenau, *Ibid.*, **38**, 747 (1931).

EFFECT OF NUCLEAR SPIN ON THE RADIATION EXCITED BY ELECTRON IMPACT

BY W. G. PENNEY*

DEPARTMENT OF PHYSICS, UNIVERSITY OF WISCONSIN

Communicated February 13, 1932

Although the polarisation of resonance radiation of components in nuclear spin multiplets has been worked out quite satisfactorily, the treatment is not so complete when the method of excitation is electron impact. It seemed worth while to investigate this point in greater detail. In the case of resonance radiation Ellett,[1] Ellett and MacNair,[2] and Larrick and Heydenburg[3] have worked out the particular case of the resonance line 2537 of mercury, the calculations allowing for the different nuclear spins of the different isotopes. The general method of attack is very simple in principle: the probability of an atom being raised from the ground state by the incident radiation to a certain Zeeman component of a higher level is found. The probability that the atom then falls to a Zeeman component of a lower level is obtained and from these two it is simple to find the polarisation of the radiation emitted. A similar method can be used if the atom is excited by electron impact. There is, however, a considerable complication introduced by the possibility of an atomic and an impinging electron changing over, with or without a change in the total spin of the atom.

The simplest case to work out would be that of an atom with one valence electron, whose nucleus had a spin $^1/_2$, for example H, or even with a general nuclear spin i, as would apply to the alkali metals. A further simplification would be to assume that the ground state of the atom had $J = 0$, as then $f = i$ for this level. For experiments on electron impact mercury, helium, zinc and cadmium are especially suitable and these are all two electron systems with a ground state $J = 0$. The actual experimental material is rather meager, sodium and mercury being the only two metals on which any data are available. That the sodium D lines are nearly or completely unpolarised has been shown by several observers.[4] For mercury 2537 there is, however, a definite polarisation depending very much on the electron velocity. Since the nuclear spin of sodium is not known we consider chiefly the case of mercury.

The question of the polarisation of the radiation excited by electron impact has been considered by Oppenheimer[5] and it was claimed in his paper that the methods of the quantum theory give a good account of all the experimental data. However, as pointed out by the same author,[6] the proof that the polarisation tends to zero as the exciting energy is decreased to the excitation potential is not valid because electron inter-

change is neglected. The rest of the paper of Oppenheimer is mainly qualitative and on pursuing the calculations further the results are disappointing. The present paper is concerned with showing that no theory of polarisation of radiation excited by electron impact can be complete without taking into account nuclear spin and exact expressions are given in a few cases. An endeavor is made to make calculations for mercury 2537, using the first order cross-section and approximate wave functions. The results do not agree at all with experimental data in the region where any measurements have been made. The conclusion is that first order cross-sections simply will not do for this type of calculation, although possibly not such a bad approximation at energies of one or two hundred volts. This conclusion is not surprising but it is interesting to have a case where first order cross-sections fail so badly, since as far as the total cross-section is concerned the results with first order cross-sections are surprisingly good.

The Atomic Wave Functions.—The type of coupling we assume is that L is coupled to S to give a resultant J which is then coupled to i to give a resultant f. The values f can assume are then $|J+i| \ldots |J-i|$. First of all we neglect the nuclear spin entirely and adopt for the atom the system of quantisation (L, S, M_L, M_S), that is, a system appropriate to the case when a strong field is acting on the atom. By introducing the orbit-spin interaction $(\mathbf{L.S})$ and making the field strength tend to zero, we get a system of quantisation (L, J, M_J) and can express the wave functions as

$$\psi(L, J, M_J) = \sum_{M_S} S^{LS}_{JM_LM_S} \psi(L, S, M_L, M_S), \qquad (1)$$

where the coefficients S are given by Wigner.[7] For two electron systems, when the singlet levels are close to the triplets, it is necessary to consider all four levels as forming one group and then the easiest way of finding the coefficients S is that of Houston.[8] For Hg, Cd, Zn, it is necessary to make this refinement or else the intersystem lines are not possible and we are very much interested in these.

Introducing the nuclear spin i, we can use a system of quantisation (L, J, M_J, i, M_i), which would be suitable if there were a weak magnetic field acting on the atom, strong enough to break down the coupling between J and i, but too weak to affect that between L and S. The wave functions for this system of quantisation will be

$$\psi(L, J, M_J, i, M_i) = \psi(L, J, M_J) e^{iM_i\phi} I(M_i), \qquad (2)$$

where $I(M_i)$ is the spin function for the nucleus in the state M_i. Introducing the coupling $(\mathbf{J}.\mathbf{i})$ and making the field strength tend to zero, we arrive at the system of quantisation needed in our problem (L, J, f, M_f) and the wave functions are

$$\psi(L, J, f, M_f) = \sum_{M_i} S^{Ji}_{fM_JM_i} \psi(L, J, i, M_J, M_i), \qquad (3)$$

where $S^{Ji}_{fM_JM_i}$ is obtained from (2) by making the suitable changes in the symbols. Equation (3) is particularly simple for states with $J = 0$. Then

$$\psi(L, J, f, M_f) = \psi(L, O, i, O, M_i). \qquad (4)$$

Excitation by Impact.—It is necessary now to calculate the probability that an electron impinging on the atom should excite it from one of the Zeeman components (f', M'_f) of the ground state 1^1S_0 to one of the Zeeman components (f, M_f) of a higher energy level. In the case of resonance radiation it is necessary to consider the intensity distribution within the line itself. This is quite unnecessary with electron impact since here there is nothing corresponding to resonance, except just at the excitation potential, where the thermal spread of the electrons is quite enough to give uniform excitation over all the hyperfine components. By considering the incident and scattered electrons as plane waves (a good approximation for energies more than 100 volts and a fair approximation at lower energies, not too small) and not considering nuclear spin at all, the author[9] has shown that the chance of exciting the Zeeman component M_J of any excited level J is given by

$$A_{M_J} = \int_0^\pi \sin\delta d\delta\, k'/k[\,|\,a_1(f_{M_J} - G_{M_J}) + a_3 g_{M_J}\,|^2 + |\,a_2 g_{M_J-1}\,|^2 + |\,a_4 g_{M_J+1}\,|^2], \qquad (5)$$

where $kh/2\pi m$ and $k'h/2\pi m$ are the velocities of the incident and scattered electrons, δ being the angle of scattering. The coefficients a are the transformation coefficients in the special case when singlets and triplets must be considered together, and the formulae for them are given by Houston.[8] f, G and g are complicated definite integrals involving the electron velocity, but whose explicit forms are not needed here. G_0 and g_0 are by far the most important of the G_M and g_M integrals, but even these are negligible at energies of 30 or 40 volts upward. At the excitation potential $f_{\pm 1}$ vanish but f_0 is quite large. As the energy increases $f_{\pm 1}$ increase and f_0 diminishes, until for very fast electrons (say 1000 volts) f_0 is inappreciable compared with $f_{\pm 1}$.

The effect of the introduction of a nuclear spin will be to modify this formula. From equations (3), (4) and (5) we see

$$A_{f'M_f'fM_f} = A_{iM_i'fM_f} = |\,S^{Ji}_{fM_JM_i'}|^2 A_{M_J},$$

since $f = i$ in the ground state. Here $M_J = M_f - M'_i$. Summing over all the excitations to the level (f, M_f) and remembering that $A_1 = A_{-1}$, we find that the total probability of this level being excited is

$$A_{fM_f} = A_0 |\, S^{ji}_{fOM_f} |^2 + A_1[\, |\, S^{ji}_{f1M_f-1} |^2 + |\, S^{ji}_{f-1\,M_f+1} |^2\,].$$

Polarisation of Radiation.—We have now to calculate the probability that an atom in an excited state (f, M_f) falls to a lower level (f', M'_f). For the parallel components we have $M'_f = M_f$ and for the perpendicular $M'_f = M_f \pm 1$. Knowing the number in the excited state and the relative intensities of the Zeeman components arising from this level, the product of the two gives the contribution to the parallel and perpendicular components from this (f, M_f). By summing over M_f we can obtain the polarisation of any hyperfine line and by summing again over f and over the different isotopes we get the polarisation of the unresolved line as it is usually observed. The relative intensities of the Zeeman components follow immediately from the usual Hönl formulae.[10] Using these and the populations in the excited states, as already explained, we get immediately the polarisation

$$P = (I_1 - I_2)/(I_1 + I_2),$$

where I_1 is the intensity parallel and I_2 is the intensity perpendicular to the electron stream, when the radiation is observed at right angles to the electron beam. We find the following expressions for the polarisation of the radiation emitted in the jump $J = 1$ to $J = 0$ and f to f'. The results are summarized in the table. Here λ and μ denote the probability of exciting the components $M_J = 0$ and $M_J = \pm 1$, respectively, in the case where there is no nuclear spin. A_1 and $A_{\pm 1}$ of equation (5) are thus first approximations to λ and μ.

TABLE

i	$f \longrightarrow f'$		P
0	1	0	$(\lambda - 2\mu)/(\lambda + 2\mu)$
$^3/_2$	$^5/_2$	$^3/_2$	$(\lambda - 26\mu)/25(\lambda + 2\mu)$
	$^3/_2$	$^3/_2$	$(7\lambda - 82\mu)/75(\lambda + 2\mu)$
	$^1/_2$	$^3/_2$	$-\,^1/_3$
$^1/_2$	$^3/_2$	$^1/_2$	$(\lambda - 4\mu)/3(\lambda + 2\mu)$
	$^1/_2$	$^1/_2$	$-\,^1/_3$

On summing over f for each value of i, we find the polarisation for $i = {}^3/_2$.

$$P = -(\lambda + 224\mu)/225(\lambda + 2\mu)$$

and for $i = {}^1/_2$

$$P = (\lambda - 10\mu)/9(\lambda + 2\mu).$$

Discussion.—For the singlet states, where the wave functions with very

good approximation are antisymmetrical in the spins, μ is zero at the excitation potential. Hence the introduction of a nuclear spin has actually lowered the polarisation (per cent) from 100 for $i = 0$ to 11 for $i = {}^1/_2$ and to $-{}^1/_4$ for $i = {}^3/_2$. As the excitation energy increases, μ increases and for each isotope the polarisation decreases and hence that of the unresolved line also. For very high velocities λ vanishes compared with μ and again the polarisation, instead of being -100 is -56 and -50, respectively. Our results extend the rules given by Skinner.[11] The rule that at the excitation potential only those states are excited for which $\Delta M_J = 0$, and the polarisation is therefore 100 per cent, is seen to require severe modification. It applies only when there is no nuclear spin and also only to singlet states, where there is no possibility of an electron interchange, with a change of spin of the atom. Again, at high velocities the rule $\Delta M_J = \pm 1$, and the polarisation is -100, applies only to cases where $i = 0$ (not merely to singlet levels this time because interchange is unimportant at high velocities).

Explanation is needed in the case of intersystem lines and an actual example will make the situation clear. With mercury, the line $1^1S_0 - 2^3P_1$ (λ 2537) is possible because the 2^3P_1 wave function is built up of parts, one antisymmetrical in the spins and the rest symmetrical. In other words, the (**L.S**) coupling has introduced a term antisymmetrical in the spins into the wave function, and it is this part which enables combinations with the ground state to occur. When the atom is excited by electrons with energy less than about 30 volts, the terms which are antisymmetrical in the orbits also contribute on account of interchange, most often with the spin of the emerging electron opposite to that of the incident. Since the antisymmetrical terms are much the larger in the 2^3P_1 wave function, interchange plays a very important part here. A small error in estimating the interchange term leads to a large error in the polarisation and this is probably the reason why our calculations (described later) are so much in error. It does seem remarkable that 2537 is unpolarised at the excitation potential because there are so many features involved which have appreciable effects. It is clear that the effect of the nuclear spin is to diminish the polarisation and the greater the nuclear spin the more complete is the depolarisation. Perhaps the result that the D lines of sodium are unpolarised when excited by electron impact can be attributed to the large nuclear spin, the exact magnitude of which, however, is not known.

It may be mentioned that a theoretical interpretation of the polarisation of the mercury lines in the visible spectrum, which has been observed quite accurately,[11] is impossible at present on account of the cascade effect. One would expect that this objection is not very serious when applied to the 1850 line, since it is so intense that atoms passing through the 2^1P_1 state on their way down to the ground state are relatively few

in number compared with those excited directly. However, the difficulties of working in this region make measurements on this line almost impossible.

Calculations on 2537.—In the hope that at least a rough check might be obtained on the theory given here, the first order cross-sections in the case of mercury 2537 were estimated and the polarisation calculated as a function of the electron energy. The calculations have been described in a previous paper.[9] The results are most disappointing and are so far from reality that they are only briefly described. The first number is the polarisation per cent for $i = 0$, the second for $i = {}^1/_2$ and the third for $i = {}^3/_2$. The number in brackets is for the unresolved line correcting for the relative abundance and different nuclear spins of the isotopes as given by Schuler and Keyston.[12] At the excitation potential the polarisation is -92, -53, -48, (-80); it rises rapidly to -34, -33, -33, (-34) at 20 volts and then falls slowly to -54, -40, -39, (-50) at 100 volts, becoming more and more negative with increasing energy. According to various observers,[11] there is no polarisation at the excitation potential but it falls steeply to -30 at 8 volts, rises to zero and is 10 at 20 volts (where the cascade effects may be considerable). At high velocities it probably becomes more and more negative. The disagreement is striking and shows that care is needed in using first order cross-sections. It seems quite certain that first order cross-sections will not do for calculations on polarisation, although it must be admitted that we have chosen about the worst possible case, where the exchange integrals have a much magnified importance.

In conclusion I wish to express my thanks to Prof. J. H. Van Vleck for many useful suggestions during the course of this work.

Summary.—It is shown that no theory of the polarisation of radiation excited by electron impact can be complete without taking into account nuclear spin. Exact expressions are given in a few special cases and applied to the particular case of mercury. The anomalous behavior of 2537 is shown to be due to the exaggerated importance which electron exchange has in exciting the 2^3P_1 level. An attempt is made to calculate the polarisation using first order cross-sections but these are not accurate enough for this purpose.

* Commonwealth Fund Fellow.

[1] A. Ellett, *Phys. Rev.*, **35**, 588 (1930).

[2] A. Ellett and W. A. MacNair, *Phys. Rev.*, **31**, 180 (1928).

[3] L. Larrick and N. P. Heydenburg, *Phys. Rev.*, **39**, 289 (1932).

[4] W. Kossel and C. Gerthsen, *Ann. d. Phys.*, **77**, 273 (1925); A. Ellett, P. D. Foote and F. L. Mohler, *Phys. Rev.*, **27**, 31 (1926).

[5] J. R. Oppenheimer, *Proc. Nat. Acad. Sci.*, **13**, 800 (1927).

[6] J. R. Oppenheimer, *Phys. Rev.*, **32**, 361 (1928).

[7] E. Wigner, *Gruppentheorie*, p. 206.

[8] W. H. Houston, *Phys. Rev.*, **33**, 297 (1928).

[9] W. G. Penney, *Phys. Rev.*, **39**, 467 (1932).

[10] See for example, H. Weyl, *Gruppentheorie*, p. 159; H. Hönl, *Zeit. f. Phys.*, **31**, 340 (1925).

[11] H. W. B. Skinner, *Proc. Roy. Soc.*, **A112**, 642 (1926); H. W. B. Skinner and E. T. S. Appleyard, *Proc. Roy. Soc.*, **A117**, 224 (1927); B. Quarder, *Zeit. f. Phys.*, **41**, 674 (1927); A. Ellett, P. D. Foote and F. L. Mohler, *Phys. Rev.*, **27**, 31 (1926); J. A. Eldridge and H. F. Olson, *Phys. Rev.* **28**, 1151 (1926).

[12] H. Schüler and J. E. Keyston, *Zeit. f. Phys.*, **72**, 423 (1931).

ENTROPY, REVERSIBLE PROCESSES AND THERMO-COUPLES

BY E. H. KENNARD

DEPARTMENT OF PHYSICS, CORNELL UNIVERSITY

Communicated February 9, 1932

It is frequently stated in the books that reversible processes never change the entropy of the universe. In justifying this statement, however, attention usually centers upon examples and in consequence a certain refinement of the statement has been overlooked which is sometimes important. The purpose of the present note is to call attention to this restriction and to discuss two examples of its application.

A complete formal proof from thermodynamic principles would presumably follow the usual course of showing that, if a reversible process could alter the universal entropy, then we could continuously convert heat into work without compensation, in violation of the Second Law. As an alternative, a proof from Statistical Mechanics would proceed by a consideration of systems in equilibrium. In either case the assumption would have to be made that the reversible process in question is independent or isolable, that is, that it can be caused to occur as nearly as desired in the complete absence of any other thermodynamically significant process. In view of this inevitable assumption an accurate statement of the theorem is obviously the following:

An *isolable* reversible process never changes the entropy of the universe.

Of the two examples to be considered, one is of practical importance, while the other is of interest because it is a case in which the unrestricted theorem is certainly not valid.

The first example has to do with thermo-electricity. Treating the Peltier and Thomson heats as reversible processes and ignoring the irreversible Ohmic heat and also thermal conduction, Kelvin deduced long ago the equation

$$\pi = T\,dE/dT \qquad (1)$$

(π = Peltier heat, T = absolute temperature, E = electromotive force.) He had misgivings about it, however, and it has indeed long been known that thermal conduction and Ohmic heat cannot simultaneously be reduced *ad libitum:* we can make the thermal conduction negligible by using very long and slim wires, or we can minimize the Ohmic heat by using short and fat ones, but we cannot do both things at once. All that can be deduced *with rigor* from thermodynamics is the Boltzmann equation:[1]

$$T\,dE/dT - \pi = 2T^{1/2}[(K_1R_1)^{1/2} + (K_2R_2)^{1/2}], \qquad (2)$$

where K and R denote total thermal conductivity and electrical resistance, respectively, between the two junctions, the subscripts referring to the two branches of the thermo-couple. In a new proof[2] recently offered by Bridgman direct use is made of the principle that reversible processes do not alter the total entropy;[3] but the principle is not really applicable here because, owing to the insuppressibility of the irreversible phenomena, the reversible processes are not isolable. Bridgman's proof thus appears to be, from the logical standpoint, no more rigorous than was Kelvin's.

In spite of its correctness in logic, however, to many physicists this objection to the theoretical proofs does not seem to be a very serious one because the two irreversible accompaniments, Ohmic heat and thermal conduction, can after all be suppressed one at a time, and it is far from obvious that they really have anything to do either with each other or with the thermoelectric phenomena. Viewed in this way, the difficulty has the appearance of a mere accident. That it is on the contrary a natural and essential one, seems to the writer, in the light of the electron theory, to be strongly suggested by our second example, having to do with thermal transpiration.

Suppose we have perfect gas in two vessels maintained from the outside at temperatures T_1 and T_2, respectively, and connected through a porous plug. Then under equilibrium conditions it is known experimentally that the pressures p_1 and p_2 in the two vessels will be connected by the equation

$$\frac{p_1}{p_2} = \left(\frac{T_1}{T_2}\right)^{1/2}. \qquad (3)$$

Now by means of a pump let us circulate a gram of gas reversibly in the usual slow way, without appreciable disturbance of the equilibrium, removing the gas from the vessel at T_1, lowering its temperature to T_2 and its pressure to p_2, and then restoring it to the vessel at T_2, after which it will flow automatically back through the porous plug into the first vessel. Let us calculate the total change in entropy arising from this "reversible process."

We may list the steps as follows, indicating for each the external work

W done by the gas and the heat H and entropy S abstracted by it from the surroundings (C_v = specific heat at constant volume, R = gas constant):

(1) remove gas from the first vessel:

$$W_1 = RT_1, \quad H_1 = 0, \quad S_1 = 0;$$

(2) expand the removed gas at T_1 from p_1 to p_2:

$$H_2 = W_2 = RT_1 \log p_1/p_2, \quad S_2 = H_2/T_1;$$

(3) cool the gas at p_2 to T_2:

$$W_3 = R(T_2 - T_1), \quad H_3 = W_3 + C_v(T_2 - T_1),$$

$$S_3 = \int_{T_1}^{T_2} (dH_3/dT)/T \, dT = (R + C_v) \log T_2/T_1;$$

(4) return the gas to the second vessel: $W_4 = -RT_2, H_4 = 0, S_4 = 0$;

(5) allow the gas to flow automatically through the plug into the first vessel. Here $W_5 = 0$, but by conservation of energy the gas must absorb heat equal to the difference of its pV works at the two ends of the plug plus $C_v(T_1 - T_2)$, or, in detail, the absorption for each rise dT in temperature as it passes along the plug is $d(RT) + C_v dT$. (The pV works done by the gas in the plug on the gas in the vessels supplies the external work in steps 1 and 4 and so keeps constant the energy of the gas in the vessels themselves.) Hence

$$W_5 = 0, \; H_5 = (R + C_v)(T_1 - T_2), \; S_5 = \int dH_5/T = (R + C_v) \log T_1/T_2.$$

The total loss of entropy by the surroundings due to our cyclic transfer of gas is thus

$$S = \Sigma S_n = S_2 = R \log p_1/p_2.$$

The gas itself, however, has the same entropy in the end as at the beginning. The entropy of the universe has thus been changed; if $T_1 > T_2$, it has actually been decreased; and this by a reversible process!

Of course it is impossible that the entropy of the universe should really be decreased on the whole by our device; as a matter of fact, if we wish to find the total change in this entropy we have yet to take account of the increase arising from the conduction of heat from one vessel to the other through the gas in the plug. It is easily shown from kinetic theory (the details will be found elsewhere[4]) that the irreversible gain in entropy due to this cause is always ample to offset any reversible decrease that can be brought about by circulating the gas with the pump.

We have thus before us a reversible process which, considered by itself, changes the entropy of the universe and may even lower it; but this re-

versible process has an irreversible accompaniment which cannot be suppressed, and the entropy is thereby saved from ever decreasing on the whole. The necessity for the restriction upon the general theorem above is thus made very evident.

There is obviously a certain analogy between this device and the thermoelectric circuit. The work done by a gram of gas in going round corresponds to the e. m. f., the heat H_2 absorbed during the reversible change of pressure to the Peltier heat, the absorption of heat in the plug to the Thomson heat. The "Peltier heat at the second junction" vanishes, to be sure, since $H_4 = 0$, but this could easily be remedied if desired by making the first expansion at T_1 to some third pressure p_3 and then, after cooling to T_2, making a second expansion or compression to p_2.

One might even apply Kelvin's argument to such a "transpiration cycle" in the same way as is done for the thermoelectric circuit. The argument amounts to making $T_1 - T_2$ and W indefinitely small and then writing for the total work, by the laws of cycles, $W = (T_1 - T_2)H/T_1$, H being the heat reversibly absorbed; from which, $H = TdW/dT$. Putting $H = \pi$, $W = E$, we have then Kelvin's equation, (1); on the other hand, in the gas-cycle case $H = H_2$ (to the first order) and the result is:

$$H_2 = TdW/dT.$$

But since $W = \sum W_n = RT_1 \log p_1/p_2 = H_2$, $dW/dT = H_2/T + RT_1 d[\log (p_1/p_2)]/dT_1 = H_2/T + R/2$ by (3), and the experimentally correct equation is:

$$H_2 = TdW/dT - R/2.$$

Thus Kelvin's (or also Bridgman's) argument as applied to the transpiration circuit, to which it seems equally applicable, leads definitely to wrong results.

The analogy between the two cases extends to the irreversible phenomena as well. To make it closest, in order to imitate the rôle of the electric field in driving the current, we may drive our gas through the plug by introducing a gravitational field without disturbing the pressures, for instance, simply by tipping the apparatus from a horizontal position toward the vertical. Then when the gas flows downward through the plug there will be an "Ohmic" evolution of heat in it due to gaseous friction which can be written Ri^2 (i = grams of gas circulated per second). As in the electric case, we can minimize this Ohmic heat by making the plug large in cross-section and short, or we can minimize the total conduction of heat through the gas by making the plug long and slender, but both effects cannot be reduced *ad libitum* simultaneously. The maximum information that can be obtained from exact thermodynamic reasoning is again expressed by the Boltzmann inequality (2), in which now K_1 stands for the

total conduction of heat through the gas in the plug per unit of temperature-difference and R_1 for the resistance as defined above. A "gas-couple" so constructed and operated appears, in fact, to be quite indistinguishable thermodynamically from a thermo-couple. If both were concealed in boxes, with means of observing only the various heats and the work (done against the pumps in one case and upon a motor in the other), an observer would be at a loss to tell which couple was which.

The properties of such transpiration systems seem to the writer to throw a real light upon the thermodynamic properties to be expected in thermo-couples. In the gas case we have the advantage of knowing the mechanism and can see just why Kelvin's argument fails. The same molecular diffusion that gives rise to the conduction of heat, combined with the existence of a fine-grained obstacle (the plug) which impedes the mass flow of the gas and so gives rise to the "resistance," results also in the phenomenon of thermal transpiration; for the occurrence of the latter both of the factors named are necessary. Thus in this case the three phenomena are clearly inter-related and the occurrence of quantities pertaining to all three in the Boltzmann equation is quite natural.

The electron theory leads us to expect something similar in the electrical case as well. We usually suppose the current-carrying electrons to move through the lattice of the conductor in much the same way as do the molecules of gas through a porous plug, and in the presence of a temperature gradient in the conductor there seems to be every reason to expect a genuine transpiration effect to occur among the electrons, resulting in a contribution to the Thomson e. m. f. analogous to the pressure difference given by (3). For further discussion of this point, however, we shall refer to the previous paper by the author.[4] It is there shown that according to the simplest form of classical theory the effects in the two branches actually balance out in such a way that the Kelvin equation follows after all. The situation has now been changed, however, by the advent of quantum theory, and it does not seem certain that the transpiration effect upon thermal e. m. f.'s will in all cases vanish.

The question thus remains one for experimental decision. Bridgman states that the observations so far made do not definitely contradict the Kelvin equations but that the experiments themselves up to the present are not all that could be desired.

[1] L. Boltzmann, *Wien. Ber.*, **96**, 1258–1297 (1887).

[2] P. W. Bridgman, *Proc. Nat. Acad. Sci.*, **15**, 765–768 (1929).

[3] Professor Bridgman very kindly amplified his paper at this point in the course of a correspondence which has greatly assisted in clarifying the writer's ideas on this subject.

[4] Cf. E. H. Kennard, *Phys. Rev.*, **22**, 617–621 (1923).

COMMENTS ON THE NOTE BY E. H. KENNARD ON "ENTROPY, REVERSIBLE PROCESSES AND THERMO-COUPLES"

BY P. W. BRIDGMAN

JEFFERSON PHYSICAL LABORATORY, HARVARD UNIVERSITY

Communicated February 9, 1932

The matter discussed by Professor Kennard in the preceding note is of sufficient importance to justify an attempt to make somewhat plainer my own point of view and to indicate what I believe still remain the most important differences between our attitudes after a personal discussion which has now extended over several years and which I am sure both of us regard as profitable.

In the first place, I must correct the impression, justified it is true by the form of presentation of my previous paper,[1] that I claimed to have a rigorous proof of Kelvin's thermoelectric relations from the assumptions of classical thermodynamics alone. An examination of the argument there presented will show that something had to be assumed in addition to the classical thermodynamics of Kelvin and Clausius. What I assumed was the existence in nature of intrinsically irreversible processes which are accompanied by definite and characteristic entropy increases whenever they occur, irrespective of other processes which may or may not be occurring simultaneously. Examples of such intrinsically irreversible processes are heat conduction and generation of Joulean heat when an electric current flows against resistance. In particular, when an amount of heat Q flows by conduction through a temperature difference ΔT at a mean temperature T the entropy increase is $Q\Delta T/T^2$. In most situations presented by classical thermodynamics the consideration of such intrinsically irreversible processes is not necessary because by a suitable arrangement of the proportions of the system the relative effect of the irreversible processes may be made vanishingly small. In the thermoelectric case, however, this is not so, the geometrical connections of thermal conduction and Joulean heating being such that both cannot be made to vanish simultaneously in comparison with the reversible effects. As Professor Kennard has intimated, this has to many seemed a more or less fortuitous circumstance, not necessarily connected with any essential feature of the thermoelectric mechanism. I also am of this opinion, and the point of my suggestion was that by allowing the irreversible processes to occur along with the reversible process, and by ascribing to each irreversible process its characteristic entropy change and zero entropy change to the reversible process, Kelvin's relation could be at once obtained. In describing what I had done in my paper I would therefore lay the emphasis on what happens during the intrinsically irreversible parts of

the operation, an assumption somewhat novel in the form in which I used it, rather than on my assumption, in common with classical thermodynamics, that the reversible part is accompanied by no change of entropy.

To me the assumption about intrinsically irreversible processes appears most plausible, although doubtless there may be difference of opinion here. In justification it may be urged that thermal conduction and Joulean heating may both be made to occur independently of each other and with the most various accompaniment of other effects. Furthermore, by making the assumption of intrinsic irreversibility, the results of classical thermodynamics may often be obtained without suppressing the irreversible phenomena, as is so often necessary. An instructive example is afforded by an alternative treatment of the thermoelectric problem. By allowing an electric current to flow in a thermoelectric circuit of the magnitude which it spontaneously assumes and which entirely dissipates the energy input in Joulean heating, and by ascribing the characteristic entropy increases to the Joulean heat generation and to thermal conduction, and zero entropy change to the remaining reversible parts of the process, one obtains again precisely the Kelvin relations, originally obtained by disregarding the irreversible aspects.

It thus appears to me that my argument involves as its most important assumption an assumption about intrinsically irreversible processes. Professor Kennard does not see this as the most vital part of my argument, but regards rather as most vital the assumption that reversible processes cannot change the entropy of the universe. With the proper understanding of what is meant by "process" I would accept as correct the postulate that a reversible process cannot change entropy. The question here is in danger of degenerating into a matter of words. Professor Kennard insists that the process must be "isolable" if the postulate is to apply. I believe that most physicists would say that "isolable" is presupposed in their definition of "process," and that in the gas transpiration example the circulation of the gas by means of pumps cannot be regarded as a "process" in the thermodynamic sense because it cannot be separated from other thermal transfer through the plug. With such an understanding of "process" the assumption of zero entropy change is unexceptionable, as Professor Kennard would be the first to agree. I believe, however, that the definition of "process" can be somewhat extended to cases, as in the thermoelectric circuit, where it is not physically insolable, and that the assumption of zero entropy change applies when such processes are reversible. According to this point of view, a "process" is characterized by a definite additive contribution, independent of the presence of other complicating factors. A "reversible process" is a process whose effects reverse sign when the independent parameters are reversed. This may be made

more concrete by application to the thermoelectric circuit. When a current flows in a couple whose junctions are at different temperatures, it is found by experiment that heat inflow or outflow are necessary at various parts of the circuit to maintain a steady state. At every part of the circuit there are three components of the net total heat flow: first, there is thermal conduction, which is in amount exactly that which would be computed from the measured temperature gradients and the known thermal conductivity of the materials, independent of the electric current or what is happening in other parts of the circuit; secondly, there is the Joulean generation of heat which is determined only by the local electric current and the resistance of the metal, independent of the presence of the thermal current or other effects; and thirdly, there is the difference between the total heating and the sum of the two first effects, which is called the Peltier heat or the Thomson heat, depending on whether the effect occurs at a junction or in the homogeneous metal. Furthermore, the Peltier heats and the Thomson heats, thus having been recognized under simple conditions and found to be connected with definite properties of the materials, are found to lead an independent existence under other conditions. The generation of Peltier and Thomson heat are, according to my proposed extended definitions, to be called processes. What is more, they are by experiment reversible processes; this means that if only the direction of current is reversed, the Peltier and Thomson heats as defined above reverse in sign, the Joulean and conduction heats being unaltered. The experimental establishment of reversibility can, of course, be only approximate, like every other datum of experiment, but it is a cardinal assumption of the argument of my paper that such reversibility has been experimentally established with sufficient exactness and that therefore there is no accompanying entropy change.

Professor Kennard's thermal transpiration cycle does not seem to me to be a "process" even in the extended sense, and for this reason I would not consider justified the application to it of an argument like that used for the thermoelectric circuit. In the thermal transpiration case the mechanism is well understood, and it is obvious that the same gas molecules which accomplish the heat transfer which brings about the net increase of entropy also take part in the convection cycle of the gas. The result is that the transfer of kinetic energy by the gas molecules from one side of the plug to the other which constitutes thermal conduction is modified by the mass motion involved in the convection cycle, since obviously during convection the kinetic energy on one side of the plug is increasing more rapidly than when there is no convection by an amount equal to the kinetic energy of the convected molecules themselves. This means that the phenomena cannot be adequately described by ascribing a thermal conduction in the ordinary sense to the gas molecules entrapped in

the plug; it is not possible to ascribe to the gas a thermal conductivity such that the net transfer of kinetic energy is equal to the product of thermal conductivity and temperature difference independent of anything else taking place. On the other hand, it was the assumption, based on experiment, that there is such a conduction in the proper sense in the thermoelectric case. In this sense the thermal transpiration cycle is not reversible, because when the direction of flow is reversed the corresponding various thermal and mechanical effects do not all reverse sign, attributing the same contribution to thermal conduction as before reversal.

To Professor Kennard's point that if a thermoelectric circuit and a gas transpiration engine were enclosed in two boxes and measurements permitted only from the outside, no difference could be found between the two boxes, I would reply that although there might be precise parallelism between the measurable quantities for a single particular set-up, nevertheless if the conditions were sufficiently varied, as by varying the temperatures of the source and the sink, it would be possible to analyze out of the data pertaining to the thermoelectric box a part perfectly reversible in the sense above, while such analysis would not be possible for the thermal transpiration box.

My final position is, therefore, that granted the experimental assumptions, my argument is to be regarded as a natural and probably correct extension of the methods of classical thermodynamics. But it is nevertheless possible to maintain the position that the parallelism between the thermoelectric circuit and the phenomena of gas transpiration is so close that one may well doubt whether as an experimental fact there may not be small irreversible aspects of the thermoelectric circuit not yet detected, and one may attempt to find such effects by better experiments. In emphasizing this aspect of the situation I believe that Professor Kennard has made an important point.

[1] P. W. Bridgman, *Proc. Nat. Acad. Sci.* **15,** 765–768(1929).

ON THE FIELD OF VALUES OF A SQUARE MATRIX

BY FRANCIS D. MURNAGHAN

DEPARTMENT OF MATHEMATICS, THE JOHNS HOPKINS UNIVERSITY

Communicated January 22, 1932

If $\mathbf{A}$ is a square n-rowed matrix of real or complex numbers whose element in the p^{th} row and q^{th} column is denoted by a_{pq}, the composite form $(\mathbf{A}\,\mathbf{x}\,|\,\mathbf{x}) \equiv \sum_{p,q}^{n} a_{pq}\, x_q\, \bar{x}_p$ assumes real or complex values $z = \xi + i\eta$, and our present purpose is to discover the region of the complex plane covered by these values under the hypothesis that the composite form $(\mathbf{x}\,|\,\mathbf{x}) = \sum_{p=1}^{n} x_p\, \bar{x}_p$ has the value unity (i.e., in the language of real geometry the *point* $\mathbf{x}$ is supposed to lie on the *unit sphere*). Since the composite form $(\mathbf{A}\,\mathbf{x}\,|\,\mathbf{x})$ has values invariant under unitary transformations of $\mathbf{A}$ the region in question is a unitary invariant of $\mathbf{A}$. For normal matrices (i.e., matrices which can be transformed unitarily to the diagonal form) it is well known and immediately evident that the region is the interior and perimeter of the convex polygon whose corners are characteristic numbers of $\mathbf{A}$ and which contains, either in its interior or on its perimeter all characteristic numbers of $\mathbf{A}$. For not-normal matrices it is known (Toeplitz-Hausdorff) that the region is convex[1] and for a two-rowed matrix the region is, according to a remark of Wintner, an ellipse having the two characteristic numbers as foci as may be shown by direct calculation in Cartesian coördinates. This method is somewhat complicated and not easy to generalize while the method of the present note using the complex variable is quite simple and works as well, in principle, for the general case as for $n = 2$.

The range of values of a Hermitian matrix is the segment of the real axis connecting the smallest and greatest characteristic numbers. Any matrix $\mathbf{A}$ can be analyzed unambiguously in the form $\mathbf{A} = \mathbf{F} + i\mathbf{G}$, where $\mathbf{F}$ and $\mathbf{G}$ are Hermitian and hence the real part of $(\mathbf{A}\,\mathbf{x}\,|\,\mathbf{x}) = (\mathbf{F}\,\mathbf{x}\,|\,\mathbf{x})$. If r denotes the largest characteristic number of $\mathbf{F}$ the line $\xi = r$ contains one point of the boundary of the sought-for region R (if it contains more than one such point a portion of this line forms part of the boundary of R), and there are no points of R to the right of this line. If now $t = e^{i\theta}$ is a turn (i.e., a complex number of unit modulus) the field of values $\tilde{R}$ of $\tilde{\mathbf{A}} = \mathbf{A}/t$ is found by rotating R around the origin through an angle $-\theta$. Hence if we find the line $\xi = p$ "bounding" $\tilde{R}$ on the right (in the sense that the line $\xi = r$ "bounds" R on the right so that, in Minkowski's terminology, $\xi = p$ is a *line of support* of $\tilde{R}$) the line obtained by

rotating $\xi = p$ around the origin through an angle θ will "bound" R in the direction θ.

Now $\mathbf{F} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ where $a^*_{pq} = \bar{a}_{qp}$ and so $\tilde{\mathbf{F}} = \frac{1}{2}(\mathbf{A}/t + \mathbf{A}^*t)$ and as p is a characteristic number of $\tilde{F}$ we must have

$$\det[p\mathbf{E} - \tfrac{1}{2}(\mathbf{A}/t + \mathbf{A}^*t)] = 0; \quad \mathbf{E} \text{ the unit matrix.}$$

Here p is the perpendicular from the origin on the "bounding" line (the argument of the perpendicular being θ) so that the equation of the "bounding" line is $\xi \cos\theta + \eta \sin\theta = p$ or, more simply, $\frac{z}{t} + \bar{z}t = 2p$. The perpendicular p_1 from any point z_1 is given by writing $z - z_1$ for z so that $p_1 = p - \frac{1}{2}\left(\frac{z_1}{t} + \bar{z}_1 t\right)$. Let us now suppose that $\mathbf{A}$ has been transformed (according to Schur's theorem) by a unitary transformation into triangle form (so that its diagonal elements are its characteristic numbers $\lambda_1, \lambda_2, \ldots \lambda_n$); the diagonal elements of the matrix $p\mathbf{E} - \frac{1}{2}(\mathbf{A}/t + \mathbf{A}^*t)$ are the perpendiculars $(p_1, \ldots p_n)$ from the points $(\lambda_1, \ldots \lambda_n)$ on the bounding line. Denoting the canonical triangle from of $\mathbf{A}$ by

$$\mathbf{C} = \mathbf{UAU}^* = \begin{pmatrix} \lambda_1 & c_{12} & \ldots & c_{1n} \\ 0 & \lambda_2 & \ldots & c_{2n} \\ . & . & \ldots & .. \\ 0 & \ldots & 0 & \lambda_n \end{pmatrix}$$

our equation is

$$\begin{vmatrix} p_1 & c_{12}/2t & .. & .. & c_{1n}/2t \\ \bar{c}_{12}t/2 & p_2 & .. & & c_{2n}/2t \\ & .. & .. & & \\ \bar{c}_{1n}t/2 & .. & .. & & p_n \end{vmatrix} = 0,$$

where, to remove the minus signs, we have multiplied all the rows and columns of the determinant by -1. For $n = 2$ this simplifies to $p_1p_2 = \frac{1}{4}c_{12}\bar{c}_{12}$ which is the familiar equation of an ellipse with foci at the characteristic numbers (λ_1, λ_2) and square of minor axis $= c_{12}\bar{c}_{12}$. For $n = 3$ the equation is

$$p_1p_2p_3 - \tfrac{1}{4}(c_{23}\bar{c}_{23}p_1 + c_{13}\bar{c}_{13}p_2 + c_{12}\bar{c}_{12}p_3) + \tfrac{1}{8}(c_{12}c_{23}\bar{c}_{13}/t + \bar{c}_{12}\bar{c}_{23}c_{13}t) = 0,$$

It is evident that the perpendicular from the point $m_1\lambda_1 + m_2\lambda_2 + m_3\lambda_3$, where $m_1 + m_2 + m_3 = 1$ is $m_1p_1 + m_2p_2 + m_3p_3$. This enables us to rewrite our equation in the form $p_1p_2p_3 = kP_\alpha$ where P_α is the perpendicular from the point

$$\alpha = (c_{23}\bar{c}_{23}\lambda_1 + c_{13}\bar{c}_{13}\lambda_2 + c_{12}\bar{c}_{13}\lambda_3 - c_{12}c_{23}\bar{c}_{13})/(c_{23}\bar{c}_{23} + c_{13}\bar{c}_{13} + c_{12}\bar{c}_{12})$$

and $k = c_{23}\bar{c}_{23} + c_{13}\bar{c}_{13} + c_{12}\bar{c}_{12}$.

In general the curve $p_1p_2p_3 = kP_\alpha$ is a curve of class three having $(\lambda_1, \lambda_2, \lambda_3)$ as foci and the joins $(\alpha_1\lambda, \alpha\lambda_2, \alpha\lambda_3)$ as tangents (the equation may be made homogeneous by using the quadratic relation between the perpendiculars from any three non-collinear points whose vanishing gives the circular points at ∞; the three perpendiculars are line coördinates corresponding to ordinary barycentric coördinates). If α coincides with any one of the points $(\lambda_1, \lambda_2, \lambda_3)$, λ_1 say, the boundary of the region sought for is found by stretching a string around the point λ_1 and the ellipse $p_2p_3 = k$.

In general, if the equation of a plane curve is given in the form $\phi(p_1, \ldots p_n) = 0$, its real foci are obtained by writing $2p_k = \frac{z-z_k}{t} + (\bar{z} - \bar{z}_\alpha)t$ and letting $t \longrightarrow 0$ (for a tangent to the curve through a focus is of the form $z = \text{const.}$ or $\bar{z} = \text{const.}$ and in order that this may be obtainable from $2p = \frac{z}{t} + \bar{z}t$ we must either have $t = 0$ or $t = \infty$). It is clear then that in the general case the characteristic numbers of the matrix are the foci of the curve bounding the field of values of the matrix; for only the term of highest degree, i.e., $p_1p_2 \ldots . p_n$ occurs in determining the foci.

[1] F. Hausdorff, *Math. Zeits.*, **3**, 314–316 (1919).

REMARKS ON THE ERGODIC THEOREM OF BIRKHOFF

By Aurel Wintner

Department of Mathematics, The Johns Hopkins University

Communicated January 22, 1932

In a recent paper[1] Birkhoff gave the proof of an extremely general recurrence theorem in dynamics which is of fundamental importance not only for stability questions connected with Celestial Mechanics, but also for the theory of adiabatic invariants and for Statistical Mechanics.[2] The object of the present note is to discuss the relation of Birkhoff's discovery with an essentially more special result, published by the writer a few years ago in connection with the theory of certain infinite matrices.[3] These matrices represent, in the sense of Frobenius, the groups defined by the almost periodic function of Bohr for which the conditionally periodic motions of Staude and Stäckel and the "grenzperiodisch" functions of H. Bohr, also occurring in dynamics,[4] are the simplest non-trivial examples. The method is, however, valid even if the recurrence character of the motion is weaker than almost-periodic (it holds for instance for functions representable only by Fourier *integrals*) and is connected in some points with several researches of N. Wiener.[5] The results which may be obtained

in this manner do not yield, of course, the perfectly general theorem of Birkhoff inasmuch as a *certain* recurrence property, which will be designated as the Hadamard condition, is *presupposed*. On the other hand if this condition is fulfilled (as it is in the simplest concrete cases mentioned above), the results are not exactly identical with the theorem of Birkhoff. For instance, in the simplest cases where the Hadamard condition is fulfilled for all motions of the phase space the exceptional zero set not excluded by the general theorem of Birkhoff may be excluded. The ergodic system of Herglotz and Artin[6] shows that for general dynamical systems the exceptional zero set of Birkhoff may effectively exist.

Let $x(t)$ be a real function defined for all real values of t and continuous in any finite range. We suppose, for simplicity, that $x(t)$ is bounded in the infinite range $-\infty < t < +\infty$ and denote by a the lower limit, by b the upper limit of $x(t)$. Finally, we suppose that the average limit

$$\mu_v = \lim_{t=+\infty} \mu_v^t; \qquad \mu_v^t = (2t)^{-1} \int_{-t}^{t} v(\tau)\, d\tau,$$

introduced by Hadamard in his researches on Dirichlet series, exists for $v = [x(t)]^n$ where n is any positive integer. We then say that the function $x(t)$ is an H-function. On denoting by ξ an arbitrary real number and by $[x \leqq \xi]_t$ the set of the time points τ for which both conditions $|\tau| \leqq t$, $x(\tau) \leqq \xi$ are fulfilled, we put

$$\rho_x^t(\xi) = (2t)^{-1} \cdot \text{mes}\, [x \leqq \xi]_t \tag{1}$$

so that the relative frequency of $x(t)$ in the range $\xi_1 \leqq x < \xi_2$ during the time elapsed from $\tau = -t$ to $\tau = t$ is represented by $\rho_x^t(\xi_2) - \rho_x^t(\xi_1)$. We now have the following

Lemma.[7] If $x(t)$ is an H-function then the limit $\lim \rho_x^t(\xi)$ where $\lim t = +\infty$ exists save at an at most countable set of values ξ. In a more precise manner there exists a monotone function $\sigma_x(\xi)$; $-\infty < \xi < +\infty$ so that

$$\sigma_x(\xi) = \lim_{t=+\infty} \rho_x^t(\xi) \tag{2}$$

in all continuity points of the function $\sigma_x(\xi)$ which is unique under the normalizing condition $\sigma_x(\xi - 0) = \sigma_x(\xi)$, $-\infty < \xi < +\infty$. Obviously

$$\sigma_x(\xi) = 0 \text{ for } \xi < a, \text{ and } \sigma_x(\xi) = 1 \text{ for } \xi > b. \tag{3}$$

Finally we have

$$\int_{-\infty}^{+\infty} v^n\, d\sigma_x(v) = \lim_{t=+\infty} (2t)^{-1} \int_{-t}^{t} (x(\tau))^n d\tau \text{ for } n = 0, 1, 2, \ldots,$$

for instance

$$\int_{-\infty}^{+\infty} d\sigma_x(v) = 1, \quad \int_{-\infty}^{+\infty} v\, d\sigma_x(v) = \mu_x, \tag{4}$$

so that the asympotic relative frequency of $x(t)$ in the range $x \leqq \xi$ is represented by $\sigma_x(\xi) = \int_{-\infty}^{\xi} d\sigma_x(v)$, provided ξ is a continuity point of σ_x.

There is of course the possibility that the *limit* function $\sigma_x(\xi)$ is constant in a subinterval $\alpha \leqq \xi \leqq \beta$ of the range $a \leqq \xi \leqq b$ of $x(t)$ [cf. (3)]. If this is *not* the case we designate $x(t)$ as a *proper* H-function. The conditionally periodic and "grenzperiodisch" functions are proper H-functions inasmuch as all almost-periodic functions are proper H-functions.[8]

It has been shown by Bohr[9] that the limit (2) does not necessarily exist at the discontinuity points ξ of the asympotic repartition function $\sigma_x(\xi)$ even if $x = x(t)$ is an almost periodic function.

Let now

$$\xi = x(t), \ \eta = y(t), \ \zeta = z(t), \ \ldots \tag{5}$$

be any sequence of N (not necessarily proper) H-functions so that (5) represents a curve in the N-dimensional phase space

$$-\infty < \xi < +\infty, -\infty < \eta < +\infty, -\infty < \zeta < +\infty, \ldots \tag{6}$$

(for a dynamical system with f degrees of freedom we have $N = 2f$). Let $R^t(\xi, \eta, \zeta, \ldots)$ denote the relative frequency of the curve (5) in the range

$$x \leqq \xi, \ y \leqq \eta, \ z \leqq \zeta, \ \ldots \tag{7}$$

of the phase space during the time elapsing from $\tau = -t$ to $\tau = t$ so that $R^t(\xi, \eta, \zeta, \ldots) = l/2t$ where l is the amount of time which the point spends during the time interval $-t \leqq \tau \leqq t$ in the range (7). The problem solved by Birkhoff affords the introduction of space averages and space densities (cf. Levi-Civita, loc. cit.) or asymptotic time averages instead of the time averages belonging to a finite range $-t \leqq \tau \leqq t$ and it concerns therefore the existence of the limit

$$S(\xi, \eta, \zeta, \ldots) = \lim_{t=+\infty} R^t(\xi, \eta, \zeta, \ldots) \tag{8}$$

which need not exist in the discontinuity points of the asymptotic repartition function[9] $S(\xi, \eta, \zeta, \ldots)$ [or even in a set of points $(\xi, \eta, \zeta, \ldots)$ having the N-dimensional measure zero]. The Lemma solves this limit problem for $N = 1$ only.

The compound relative frequency $R^t(\xi, \eta, \zeta, \ldots)$ is not simply the product of the partial relative frequencies $\rho_x^t(\xi)$, $\rho_y^t(\eta)$, $\rho_z^t(\zeta)$, ... [defined by (1)] inasmuch as these component probabilities are not inde-

pendent of each other. However, if the functions are, for instance, almost-periodic then $R^t(\xi, \eta, \zeta, \ldots)$ may be represented[10] as a folding expression ("Faltung") of the functions $\rho_X^t(\xi)$, $\rho_Y^t(\eta)$, $\rho_Z^t(\zeta)$, ... where

$$X = x(t).\left\{\begin{matrix}\cos\\ \sin\end{matrix}\right\}(\lambda t),\ Y = y(t).\left\{\begin{matrix}\cos\\ \sin\end{matrix}\right\}(\lambda t),\ \ldots;\ -\infty < \lambda < +\infty$$

and the existence of the Birkhoff limit (8) follows then by virtue of the theorems of Helly[11] on sequences of monotone functions.

If all N-functions (5) are almost-periodic functions or, in a more general manner, proper H-functions then the asymptotic repartition function $S(\xi, \eta, \zeta, \ldots)$ cannot be constant in an N-dimensional vicinity of a point of the phase space which coincides with a point of the curve (5). In other words the limit density of probability is then in the full range of the curve (5) different from zero.

In a subsequent note the Birkhoff theorem will be applied, in an explicit manner, to the Levi-Civita-Bohl differential equation[12]

$$dx/dt = f(x, t) \text{ where } f(x, t) = f(x + 1, t) = f(x, t + 1)$$

and to connected questions in Celestial Mechanics.

[1] G. D. Birkhoff, *Wash. Proc.*, **17**, 656–660 (1931).
[2] T. Levi-Civita, *Hamb. Abh.*, **6**, 330–348 (1928).
[3] A. Wintner, *Math. Zeit.*, **30**, 310–311 (1929).
[4] G. D. Birkhoff, *Acta. Math.*, **50**, 359–379 (1927).
[5] Cf., *Ibid.*, **55**, 257–258 (1931).
[6] E. Artin, *Hamb. Abh.*, **3**, 170–175 (1927).
[7] A. Wintner, loc. cit.
[8] A. Wintner, loc. cit., p. 312.
[9] H. Bohr, *Danske Videnskab. Selsk.*, **10**, Nr. 10 (1930).
[10] For almost periodic functions cf. A. Wintner, loc. cit., p. 318–319, for functions with a more general recurrence character the researches of N. Wiener.
[11] Cf., for instance, A. Wintner, *Spektraltheorie der unendlichen Matrizen*, Leipzig, 1929, Chapter II.
[12] H. Kneser, *Math. Annalen*, **91**, 154 (1924). For the astronomical literature of the repartition problem cf. A. Wintner, *Rend. Acc. Lincei*, **11** [6], 464 (1930) and H. Bohr, loc. cit.

A GEOMETRICAL DETERMINATION OF THE CANONICAL QUADRIC OF WILCZYNSKI

BY E. B. STOUFFER

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF KANSAS

Communicated February 3, 1932

In the study of the projective differential properties of curved surfaces a canonical development for the equation of the surface and a geometrical determination of the associated tetrahedron of reference are of fundamental importance. Wilczynski[1] was the first to solve this problem. The vertices of his canonical tetrahedron can be easily located as soon as a certain quadric surface Q upon which they lie is determined. Green[2] and others[3] have obtained similar canonical developments which take their simplest forms when the vertices of the associated tetrahedrons lie upon Q.

The quadric Q is commonly called the *canonical quadric of Wilczynski*. It was located by Wilczynski by means of a unodal cubic surface osculating the curved surface. The introduction of this canonical cubic surface complicates the situation considerably, especially since its determination before Q is known is a rather involved process. Recently Bompiani[4] obtained Q—apparently without recognizing it—by a geometrical process somewhat simpler than that of Wilczynski but little related to the problem of determining a canonical tetrahedron.

In the present paper the canonical quadric of Wilczynski is defined geometrically in an exceedingly simple manner by means of the axis of Cech, a line which is covariantly related to the curved surface and which is easily located geometrically.

Since most of the preliminary facts used in this paper are found in the memoir by Green,[2] we shall use his notation in order to facilitate reference. Accordingly, we refer our curved surface S to its asymptotic net and take the associated system of differential equations in the form

$$\begin{aligned} y_{uu} + 2by_v + fy &= 0, \\ y_{vv} + 2a'y_u + gy &= 0. \end{aligned} \tag{1}$$

Let the two asymptotic curves passing through a point y of the surface be denoted by C_u and C_v. The tangents at points of C_v to the curves $v = \text{const.}$ generate a non-developable ruled surface $R^{(u)}$ and the tangents at points of C_u to the curves $u = \text{const.}$ generate a similar surface $R^{(v)}$. The points

$$\rho = y_u - \beta y, \quad \sigma = y_v - \alpha y, \tag{2}$$

where α and β are functions of u and v, lie on the tangents to C_u and C_v,

respectively, and the line l determined by them lies in the plane tangent to S at y. The tangent planes to $R^{(u)}$ at ρ and to $R^{(v)}$ at σ intersect in a line l' which passes through y and the point $y_{uv} - \alpha y_u - \beta y_v$. Green calls either of the lines l, l' the *reciprocal* of the other.

If $\beta = a'_u/2a'$, $\alpha = b_v/2b$, the lines l and l' are the directrices of Wilczynski of the first and second kind, respectively, and if $\beta = -b_u/4b$, $\alpha = -a'_v/4a'$, they are the canonical edges of Green of the first and second kind, respectively. These two pairs of lines have been located geometrically in several different ways. For example, in the former case the point ρ is the harmonic conjugate[5] of y with respect to the points where the flecnode curves of $R^{(u)}$ intersect $y\rho$, and in the latter case the point ρ is the pole of the tangent at y to C_v with respect to the osculating conic[6] of C_u at y. The corresponding points σ may be located in a similar manner.

The directrix of the first kind of Wilczynski and the canonical edge of the first kind of Green intersect in a point called the *canonical point* and their reciprocals lie in a plane called the *canonical plane*. The intersection of the canonical plane and the tangent plane to S at y is called the *canonical line*.

The harmonic conjugate of the directrix of the second kind of Wilczynski with respect to the canonical line and the canonical edge of the second kind of Green is the *projective normal* of Fubini, the pseudo-normal of Green. For it $\beta = -\frac{1}{2}(b_u/b + a'_u/a')$, $\alpha = -\frac{1}{2}(a'_v/a' + b_v/b)$. Again, the harmonic conjugate of the projective normal with respect to the directrix of Wilczynski and the canonical edge of Green, both of the second kind, is the axis of Cech. It is given by $\beta = -\frac{1}{6}(b_u/b - a'_u/a')$, $\alpha = -\frac{1}{6}(a'_v/a' - b_v/b)$. Both the projective normal and the axis of Cech have other geometrical definitions.

Green obtained a general expression representing a group of canonical developments, including that obtained by Wilczynski by choosing as the vertices of the canonical tetrahedron the four points

$$y,\ \rho = y_u - \beta y,\ \sigma = y_v - \alpha y,\ \tau = y_{uv} - \alpha y_u - \beta y_v + \alpha\beta y, \qquad (3)$$

where α and β are functions of u and v which need only be properly assigned in order to obtain any one of the several developments. Since any point X in space is defined by an expression of the form $x_1 y + x_2\rho + x_3\sigma + x_4\tau$, the coördinates of X may be taken to be (x_1, x_2, x_3, x_4). In this new coördinate system the general expression for the several canonical developments has the form

$$\begin{aligned}\zeta = \xi\eta + \tfrac{2}{3}b\xi^3 + \tfrac{2}{3}a'\eta^3 + \tfrac{1}{6}(4b\beta + b_u)\xi^4 + \tfrac{2}{3}(b_v - 2b\alpha)\xi^3\eta \\ + \tfrac{2}{3}(a'_u - 2a'\beta)\xi\eta^3 + \tfrac{1}{6}(4a'\alpha + a'_v)\eta^4 + \ldots,\end{aligned} \qquad (4)$$

where $\xi = x_2/x_1$, $\eta = x_3/x_1$, $\zeta = x_4/x_1$. Cutting off the development after the first term we have

$$x_1x_4 - x_2x_3 = 0, \tag{5}$$

which is the equation of the canonical quadric Q of Wilczynski. Equation (5) is independent of α and β and the quadric is the same for all developments (4). The lines $y\rho$, $y\sigma$, $\rho\tau$, $\sigma\tau$ lie on the quadric.

For given functions α and β the points ρ and σ generate two surfaces S_ρ and S_σ. The tangent at ρ to the curve u = const. on S_ρ and the tangent at σ to the curve v = const. on S_σ intersect the corresponding line l' in the points $\rho_v - \alpha\rho = \tau - \beta_v y$ and $\sigma_u - \beta\sigma = \tau - \alpha_u y$, respectively. The harmonic conjugate of these two points with respect to y is the point $\tau - \frac{1}{2}(\alpha_u + \beta_v)y$. This point is τ and lies on Q if α and β are chosen to satisfy the condition

$$\alpha_u + \beta_v = 0. \tag{6}$$

But this condition is satisfied if l' is the axis of Cech.

THEOREM. *Let l' denote the axis of Cech associated with a point y of a curved surface and let ρ and σ denote the points where its reciprocal intersects the tangents at y to the asymptotic curves C_u and C_v, respectively. The tangent at ρ to the curve generated by ρ for u = const. and the tangent at σ to the curve generated by σ for v = const. intersect l' in two points such that the harmonic conjugate of y with respect to them is a point τ on the canonical quadric of Wilczynski. The lines $\rho\tau$ and $\sigma\tau$ thus geometrically characterized lie on the quadric.*

Equation (4) shows that the equation of any non-singular quadric surface which has contact of the second order with S at y is of the form

$$x_1x_4 - x_2x_3 + Ax_2x_4 + Bx_3x_4 + Cx_4^2 = 0.$$

The lines $x_1 = x_2 = 0$ and $x_1 = x_3 = 0$ lie on this quadric only if $A = B = C = 0$.

THEOREM. *The canonical quadric of Wilczynski at a point y of a curved surface S is the unique quadric which has second order contact with S at y and which contains the lines $\rho\tau$ and $\sigma\tau$ associated with the axis of Cech.*

With Q thus located, any geometrically characterized pair of reciprocal lines determines by their intersections with Q a tetrahedron and a corresponding canonical development. For example, the directrices of Wilczynski give the development used by Wilczynski, the canonical edges of Green give that used by Green, and the projective normal and its reciprocal, or the axis of Cech and its reciprocal, give similar developments.

The osculating cubic surface which Wilczynski found it necessary to introduce in order to locate Q is easily determined with Q already known. It is, in fact, the cubic surface for which (0, 0, 0, 1) is a unode and $x_1 = 0$ the uniplane and which has third order contact with S at y. This cubic surface, which is of course different for each canonical development,

gives an easy method of characterizing the unit point of the coördinate system. If we impose the conditions, for example, that the coefficients of the first, second and third degree terms in (4) shall all be unity, then the unit point lies on Q and its projection from (0, 0, 0, 1) upon $x_4 = 0$ lies on a tangent of Segre. This point, the two points of intersection of the tangent with the above cubic surface, and the point of intersection of the tangent and the line $x_1 = x_4 = 0$ have a cross ratio equal to -2. The three choices of the tangent of Segre correspond to the cube roots of unity which appear when the above conditions on the coefficients are imposed.

[1] Wilczynski, *Trans. Amer. Math. Soc.*, **9**, 79–120 (1908).

[2] Green, *Ibid.*, **20**, 79–153 (1919).

[3] Stouffer and Lane, *Bull. Amer. Math. Soc.*, **34**, 460 (1928).

[4] Bompiani, *Rend. Acc. Lincei*, [6] **6**, 187–190 (1927), and *Math. Zeitschrift*, **29**, 678–683 (1929).

[5] Fubini e Cech, *Geometria Proiettiva Differenziale*, **1**, 148 (1926).

[6] Stouffer, *Bull. Amer. Math. Soc.*, **34**, 301 (1928).

DYNAMICAL SYSTEMS OF CONTINUOUS SPECTRA

BY B. O. KOOPMAN AND J. v. NEUMANN

DEPARTMENTS OF MATHEMATICS, COLUMBIA UNIVERSITY AND PRINCETON UNIVERSITY

Communicated January 21, 1932

1. In a recent paper by B. O. Koopman,[1] classical Hamiltonian mechanics is considered in connection with certain self-adjoint and unitary operators in Hilbert space $\mathfrak{H}$ $(= \mathfrak{L}_2)$. The corresponding canonical resolution of the identity $E(\lambda)$, or "spectrum of the dynamical system," is introduced, together with the conception of the spectrum revealing in its structure the mechanical properties of the system.[2] In general, $E(\lambda)$ will consist of a discontinuous part (the "point spectrum") and of a continuous part. The case of a pure point spectrum, and the other extreme, that in which the inner product $(E(\lambda)f, g)$ is, for every f and g in $\mathfrak{H}$, the Lebesgue integral of one of its derivatives, may readily be treated by known analytical tools.[3] The present paper is devoted to the case where $E(\lambda)$ is continuous $(\lambda \neq 0)$, but without $(E(\lambda)f, g)$ being necessarily equal to the integral of its derivative. It will further be assumed that the system is non-integrable in the sense that any f in $\mathfrak{H}$ such that $U_t f = f$ almost everywhere on Ω must be almost everywhere constant. In other words, we are assuming the following hypothesis:

C. $E(\lambda + 0) - E(\lambda - 0) = 0$, for $\lambda \neq 0$:

If $E_0 = E(+0) - E(-0)$, then $E_0 f =$ almost everywhere constant.

Suppose that ϕ is a characteristic function of a non-integrable system: $U_t\phi = e^{i\lambda t}\phi$. Then $U_t|\phi| = |\phi|$, so that $|\phi|$ must be almost everywhere constant, and hence we may take $\phi = e^{i\theta}$, θ being defined almost everywhere (mod 2π).[4] From $U_t\phi = e^{i\lambda t}\phi$ follows that $U_t\theta = \theta + \lambda t$ (mod 2π); thus θ is an "angle variable." Since ϕ is in $\mathfrak{H}$, $\|\phi\| = \int_\Omega e^{i\theta}\, e^{-i\theta}\, dv = \mu\Omega$ must be finite, so that a non-integrable system of infinite $\mu\Omega$ can have no characteristic function. These considerations permit the following restatement of our hypothesis:

C'. The system possesses no invariant subset of Ω of positive finite measure, and in the case where $\mu\Omega$ is finite, no angle variables.

We will show that under this hypothesis all the initially observed properties of the system are obliterated by the lapse of time: the method of elementary mechanics of computing the final from the initial state must be replaced by the methods of the theory of probability. Contrary to the case of classical statistical mechanics, this situation is not dependent upon the system's having an enormous number of degrees of freedom: this number may perfectly well reduce to two.

2. By a P-set we shall mean a set of points of Ω, by a t-set, a set of points on the time axis. The P-set for which $f(P) > a$, etc., shall be denoted as usual by $[f(P) > a]$, etc., and similarly for t-sets and functions of t. Along with the μ-measure μM of a P-set M defined with respect to the volume element $dv = \rho d\omega$, we shall define the τ-measure τI of a t-set I as follows:

$$\tau I = \lim_{T=+\infty} \frac{m(I\cdot[\,|t| \leqq T])}{2T}.$$

Here m denotes the ordinary Lebesgue measure, and the customary set-multiplication is referred to.[5] By a zero P-set or a zero t-set we shall mean one for which the μ-measure or the τ-measure, respectively, is zero. The "characteristic function" of a P-set of a t-set Θ shall be denoted, as usual, by $\chi_\Theta : \chi_\Theta(P) = 1$ or 0 according as P is or is not on Θ, etc. In these terms we are able to state the fundamental theorem of this paper:

THEOREM I. Under the hypothesis *C*, there exists a zero t-set I such that, for any two P-sets M and N of finite μ-measure, as $t \longrightarrow +\infty$ (or $t \longrightarrow -\infty$) through values not on I,

$$\mu(M_t\cdot N) \longrightarrow \frac{\mu M\cdot\mu N}{\mu\Omega}, \qquad (1)$$

M_t being the image of M after the lapse of time t. When $\mu\Omega = \infty$, the right-hand member is to be replaced by zero.

Proof.—It has been shown[6] that in non-integrable systems,

$$(E_0\chi_M, \chi_N) = \frac{\mu M \cdot \mu N}{\mu\Omega};$$

hence (1) is equivalent to

$$\lim_{\substack{t \longrightarrow \pm\infty \\ t \text{ not on } I}} (U_t\chi_M, \chi_N) = (E_0\chi_M, \chi_N),$$

which, in turn, may be replaced by

$$\lim_{\substack{t \longrightarrow \pm\infty \\ t \text{ not on } I}} (U_t f, g) = (E f, g), \tag{1'}$$

since the aggregate of χ-functions and their linear combinations is everywhere dense in $\mathfrak{H}$. That is, we must prove that as $t \longrightarrow \pm\infty$ through values not on I, $U_t f$ converges "weakly" to $E_0 f$.[7] Now $(U_t f, g) \longrightarrow (E_0 f, g)$ for all f, g if and only if $(U_t f, f) \longrightarrow (E_0 f, f)$ for all f.[8] Using, now, the known properties:[9] $U_t E_0 = E_0 U_t = U_t$, $E_0^2 = E_0$, etc., we have $((U_t - E_0)f, f) = ((U_t - U_t E_0)f, f) = (U_t(I - E_0)f, f) = (U_t(I - E_0)^2 f, f) = (U_t(I - E_0)f, (I - E_0)f)$, so that our problem is to show that $(U_t(I - E_0)f, (I - E_0)f) \longrightarrow 0$. That is, we are to show that a zero t-set I exists such that, as $t \longrightarrow \pm\infty$, t not on I, $(U_t g, g) \longrightarrow 0$ whenever $E_0 g = 0$. Since U_t is a bounded operator, it is sufficient to show this only for an everywhere dense sequence $g_1, g_2, g_3, \ldots$ of functions g.

We now introduce the expression

$$\frac{1}{2T}\int_{-T}^{T} |\, (U_t g, g)^2 \,|\, dt.$$

It has the value

$$\frac{1}{2T}\int_{-T}^{T}\left|\int_{-\infty}^{+\infty} e^{i\lambda t}\, d_\lambda \,||\, E(\lambda)g \,||^2\right|^2 dt$$

$$= \frac{1}{2T}\int_{-T}^{T} dt \int_{-\infty}^{+\infty} e^{i\lambda t}\, d_\lambda \,||\, E(\lambda)g \,||^2 \cdot \int_{-\infty}^{+\infty} e^{-i\mu t}\, d_\mu \,||\, E(\mu)g \,||^2$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(\frac{1}{2T}\int_{-T}^{T} e^{i(\lambda-\mu)t}\, dt\right) d_\lambda \,||\, E(\lambda)g \,||^2 \cdot d_\mu \,||\, E(\mu)g \,||^2$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{e^{i(\lambda-\mu)T} - e^{-i(\lambda-\mu)T}}{2i(\lambda-\mu)T} \cdot d_\lambda \,||\, E(\lambda)g \,||^2 \cdot d_\mu \,||\, E(\mu)g \,||^2$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\sin(\lambda-\mu)T}{(\lambda-\mu)T} \cdot d_\lambda \,||\, E(\lambda)g \,||^2 \cdot d_\mu \,||\, E(\mu)g \,||^2.$$

Since $E(\lambda)$ has no point-spectrum except possibly for $\lambda = 0$, for which value precisely we have $(E(+0) - E(-0))g = E_0 g = 0$, $||\, E(\lambda) \,||^2$ is a

continuous function of λ, which, as λ goes from $-\infty$ to $+\infty$, goes in a non-decreasing manner from 0 to $\| g \|^2$. If $\lambda = \phi(x)$ is the inverse of $x = \| E(\lambda) \|^2$, $\phi(x)$ increases monotonically from $-\infty$ to $+\infty$ as x goes from 0 to $\| g \|^2$, it is constant on no interval, although it may have finite jumps. Hence the above expression may be written as

$$\int_0^{\| g \|^2} \int_0^{\| g \|^2} \frac{\sin\{\phi(x) - \phi(y)\}T}{\{\phi(x) - \phi(y)\}T}\, dx\, dy.$$

Since the region of integration is finite and the integrand in absolute value less than 1, and since (except for the zero set $x = y$) it approaches zero, the expression approaches zero as $T \longrightarrow \pm\infty$. (This would not be true if $\phi(x)$ were constant on certain intervals, i.e., if $\| E(\lambda)g \|^2$ had discontinuities, and thus, $E(\lambda)$ had a point spectrum.)

Thus,

$$\lim_{T \longrightarrow +\infty} \frac{1}{2T} \int_{-T}^{+T} |(U_t g, g)|^2\, dt = 0. \tag{2}$$

Consider the series

$$\pi(t) = \sum_{n=1}^{\infty} \frac{1}{2^n \| g_n \|^2} |(U_t g_n, g_n)|^2.$$

It is uniformly convergent for $|t| < +\infty$, since it is dominated by the series $\sum_{n=1}^{\infty} \frac{1}{2^n}$, on account of the inequality $\|(U_t g_n, g_n)\|^2 \leqq \| U_t g \| \cdot \| g \| = \| g \|^2$. Hence, in virtue of (2),

$$\lim_{t \longrightarrow +\infty} \frac{1}{2T} \int_{-T}^{T} \pi(t)\, dt = 0. \tag{3}$$

Now if we can show that a zero t-set I can be found such that

$$\lim_{\substack{t \longrightarrow \pm\infty \\ t \text{ not on } I}} \pi(t) = 0, \tag{4}$$

then, on account of the inequality

$$0 \leqq |(U_t g_n, g_n)|^2 \leqq 2^n \| g \|^2 \pi(t),$$

we shall have the consequence that

$$\lim_{\substack{t \longrightarrow \pm\infty \\ t \text{ not on } I}} (U_t g_n, g_n) = 0,$$

that is, the theorem will be proved.

Let $I_m = \left[\pi(t) \geqq \frac{1}{m}\right]$. Evidently $I_1 \subset I_2 \subset I_3 \subset \ldots$, and, in virtue of (3),

$$\lim_{T \to +\infty} \frac{m(I_m \cdot [|t| \leqq T])}{2T} = 0.$$

We now choose $T_1 < T_2 < \ldots$ such that the left-hand member of the above is, for all $m = 2, 3, \ldots$, and $T \geqq T_{m-1}$, less than $\frac{1}{m}$. Then we set

$$I = I_1 \cdot [|t| \leqq T_1] + I_2 \cdot [T_1 < |t| \leqq T_2] + \ldots$$
$$\ldots + I_m \cdot [T_{m-1} < |t| \leqq T_m] + \ldots.$$

For $T_{m-1} < T \leqq T_m$ we have

$$I \cdot [|t| \leqq T] \subset I_{m-1} \cdot [|t| \leqq T_{m-1}] + I_m \cdot [|t| \leqq T],$$

hence,

$$\frac{m(I \cdot [|t| \leqq T])}{2T} \leqq \frac{m(I_{m-1} \cdot [|t| \leqq T_{m-1}]) + m(I_m \cdot [|t| \leqq T])}{2T}$$
$$\leqq \frac{m(I_{m-1} \cdot [|t| \leqq T_{m-1}])}{2T_{m-1}} + \frac{m(I_m \cdot [|t| \leqq T])}{2T}$$
$$\leqq \frac{1}{m-1} + \frac{1}{m}.$$

As $m \longrightarrow +\infty$, that is, as $T \longrightarrow +\infty$, this approaches zero; hence I is a zero t-set. Since, further, when $|t| > T_{m-1}$, a t on I_m is on I, it follows that outside I, $\pi(t) < \frac{1}{m}$. Hence as $m \longrightarrow +\infty$, or $t \longrightarrow \pm\infty$, $\pi(t) \longrightarrow 0$, so that (4) is established.

Corollary.—The set I of Theorem I can be taken as a set of intervals such that there are only a finite number of intervals of the set in any arbitrarily chosen finite interval.

Proof.—Since the expressions $(U_t g, g)$ are continuous function of t, $\pi(t)$ is likewise, so that every I_n is a closed set, and, likewise, I. Hence the Lebesgue and that Jordan measure of I are equal. In each of the regions $|t| \leqq 1$, $|t| \leqq 2$, ..., $m-1 < |t| \leqq m$, ... we can replace I by a finite set of overlapping intervals without increasing its measure by more than, for example, $\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^m}, \ldots$,—thus, in all, by a quantity < 1, so that it remains a zero t-set.

THEOREM II. When the hypothesis C is not realized, the conclusion of Theorem I is invalid.

Proof.—To begin with, suppose the system to be integrable. Then there will exist an invariant $\Lambda \subset \Omega : \Lambda_t = \Lambda$, with $\mu\Lambda > 0$ and finite, and

$\mu(\Omega - \Lambda) > 0$. If $\mu\Omega$ is finite, take $M = \Lambda$, $N = \Omega - \Lambda$ in (1), when $\mu(M_t \cdot N) = \mu(\Lambda_t \cdot (\Omega - \Lambda)) = \mu(\Lambda \cdot (\Omega - \Lambda)) = 0$, whereas $\frac{\mu\Lambda \cdot \mu(\Omega - \Lambda)}{\mu\Omega} \neq 0$. If $\mu\Omega = \infty$, take $M = N = \Lambda$, in which case $\mu(M_t \cdot N) = \mu(\Lambda_t \cdot \Lambda) = \mu\Lambda > 0$, whereas $\frac{\mu\Lambda \cdot \mu\Lambda}{\mu\Omega} = 0$. Thus, in either case, (1) is violated.

Suppose, on the other hand, that the system is non-integrable, but that it has a characteristic function φ of characteristic number $\lambda > 0$: $U_t\varphi = e^{i\lambda t}\varphi$. In this case, as we have seen, we may take $\varphi = e^{i\theta}$, and $\mu\Omega$ will be finite. In (1′) take $f = g = \varphi$; then $(U_t f, g) = (e^{i\lambda t}\varphi, \varphi) = e^{i\lambda t}(\varphi, \varphi) = e^{i\lambda t}\mu\Omega$, so that (1′), which in the non-integrable case is a consequence of (1), is impossible.

Theorem II is thus proved.

Now let us consider to what extent the presence of the exceptional t-set I appears necessary in Theorem I. We have:

THEOREM III. There exist spectra $E(\lambda)$ (not necessarily belonging to a dynamical system[10]) such that, under the hypothesis C, the conclusion (1′) of Theorem I fails to hold when I is empty.

Proof.—Take f such that $E_0 f = 0$; then it is sufficient to show that $\lim_{t \to \pm\infty} (U_t f, f)$ does not exist, or else that it is not zero. We have[11]

$$(U_t f, f) = \int_{-\infty}^{+\infty} e^{i\lambda t} d \,||\, E(\lambda) f \,||^2,$$

and since $x = ||\, E(\lambda) f \,||^2$ can evidently be taken to be an arbitrarily given continuous non-decreasing function, increasing from zero to a finite a $(= ||\, f \,||^2)$ as t goes from $-\infty$ to $+\infty$, its inverse $\lambda = \varphi(x)$ can obviously be taken to be equal to an arbitrary monotonically increasing function, possibly with finite jumps, in the interval of definition $0 \leqq x \leqq a$, and going from $-\infty$ to $+\infty$. Hence

$$(U_t f, f) = \int_0^a e^{i\varphi(x)t}\, dx.$$

Now let $a = 1$, and $\varphi(x)$ have always such a value that its hexadic development contains only the digits 0 and 5. For instance, let $\varphi(0) = -\infty$, $\varphi(1) = +\infty$, and, for $x = \sum_{n=1}^{\infty} \frac{a_n}{2^n}$ ($a_n = 0$ or 1, the latter infinitely often), let $\varphi(x) = \sum_{n=1}^{\infty} \frac{5a_n}{6_n}$. For $t = 2\pi \cdot 6^n$, the fractional part of $\frac{1}{2\pi}\ \varphi(x) \cdot t$ will thus start with the hexadic digit 0 or 5, i.e., it is $\geqq 0$, $\leqq \frac{1}{6}$, or $\geqq \frac{5}{6}$,

$\leqq 1$—i.e., $\frac{1}{2\pi}\varphi(x)t$ has a value (mod 1) $\geqq -\frac{1}{6}, \leqq \frac{1}{6}$. Hence

$$\mathfrak{R}\int_0^a e^{i\varphi(x)t}\,dx = \int_0^1 \cos\left(2\pi\cdot\frac{1}{2\pi}\varphi(x)t\right)dx$$

$$\geqq \int_0^1 \cos\left(2\pi\cdot\frac{1}{6}\right)dx$$

$$= \cos\frac{\pi}{3} = \frac{1}{2} > 0.$$

Thus $\lim_{t\to\pm\infty}(U_t f, f) = 0$ is excluded, and our theorem is proved.

III. Theorems I, II, III, delimit a situation which could scarcely be rendered more precise: the hypotheses and diverse restrictions seem all inherent in the nature of the case. It would, of course, be extremely interesting if a proof of Theorem I could be found without the use of the spectrum $E(\lambda)$, etc., for example, along the lines of the recent paper of Birkhoff's on the ergodic hypothesis.[12]

Theorem I expresses the fact that in a system for which C holds, consecutive states at two sufficiently different epochs are almost certainly statistically independent. For $\frac{\mu(M_t\cdot N)}{\mu\Omega}$ is the probability that, in the time t, the system go from M to N—and this is $= \frac{\mu M\cdot\mu N}{(\mu\Omega)^2}$, i.e., the product of the probabilities of the system's being in M and of its being in N. (We say "almost certainly," to correspond with the fact that the exceptional zero t-set I is excluded.) Such a system can have, *a la longue*, no physical properties. Indeed, the only properties which any system can have *a la longue* are those which violate C, namely:

1. Invariant sub-sets: barriers that are never passed.
2. Angle variables: clocks that never change.[13]

Theorem I may also be expressed by saying that the states of motion corresponding to any set M of Ω become more and more spread out into an amorphous everywhere dense chaos. Periodic orbits, and such like, appear only as very special possibilities of negligible probability.

We have already called attention to the essential difference between this situation, which may exist in very simple systems, and that envisaged in the kinetic theory of gases, in which the confused character of the motion is an intuitively evident consequence of the large number of degrees of freedom of the system.

[1] "Hamiltonian Systems and Transformations in Hilbert Space," these PROCEEDINGS, 315–318 (May, 1931); this reference will here be abbreviated to (H). Another reference of importance for us is the paper by v. Neumann on the proof of the quasi-ergodic hy-

pothesis (these PROCEEDINGS, pp. 70–82 (Jan., 1932)), which will be referred to under the abbreviation (Q).

We shall, in the present paper, assume the notation, etc., of (H) and (Q) to be known.

[2] Properties, of course, which are true "almost everywhere" on the manifold of states of motion Ω, corresponding with the nature of the functional tools. The allied conception of properties true "almost everywhere" on the time axis with respect to the time-average τ-measure here defined is an essential notion in the present paper; it is contained implicitly in J. v. Neumann's recent proof of the quasi-ergodic hypothesis (these PROCEEDINGS, pp. 70–82 (Jan., 1932), and 263–266 (March, 1932)).

[3] These cases are made the subject of as yet unpublished investigations by B. O. Koopman. In the case of a pure point spectrum, the theory of almost periodic functions is available: $(U_t f, g)$ is an almost periodic function of t, a fact having an obvious physical interpretation when $f = \chi_M(P)$ and $g = \chi_N(P)$, characteristic functions of the sets M and N ($\subset \Omega$; μM, μN finite). In the second case here noted, in virtue of the well-known Fourier integral theorem, $(U_t f, g) \longrightarrow 0$ as $t \longrightarrow \infty$, showing (on making the above choice of f, g) that any such region M will flow, in course of time, "almost entirely" out of a fixed region N of finite measure. This possibility obviously implies that $\mu\Omega$ is infinite. There is a similarly obvious interpretation when $\lambda = 0$ is a simple unique characteristic number.

[4] Cf. (H), p. 318.

[5] It is unnecessary for us to enter upon the discussion of τ-measurability or the properties of τ-measure.

[6] Cf. (Q), p. 78.

[7] The ergodic theorem (Q) states that $\frac{1}{T}\int_0^T U_t f\, dt$ converges "strongly" to $E_0 f$, thus this theorem tells us *more* than the present one from the point of view of convergence, but *less*, from the point of view of the function, since f is replaced by its time average.

[8] If we replace in $(U_t f, f) \longrightarrow (E_0 f, f)$ f by $\frac{f+g}{2}$ and by $\frac{f-g}{2}$, and subtract, the real part of $(U_t f, g) \longrightarrow (E_0 f, g)$ results. If we replace g by ig, the imaginary part results. Thus the general formula holds.

[9] Cf. (Q), p. 73.

[10] The following question is of great interest: When will a canonical resolution of the identity $E(\lambda)$ be the spectrum of a dynamical system? An (unpublished) formula obtained by Koopman is that, when f, g, and the ordinary product fg are in $\mathfrak{H}$, then, for a dynamical system,

$$E(x)fg = \int_{-\infty}^{+\infty} E(x - s)f \cdot d_s E(s)g,$$

which is a consequence of the fundamental equation obtained in (H):

$$U_t F(f, g, \ldots) = F(U_t f, U_t g, \ldots),$$

where F is any single-valued function. Recently J. v. Neumann has shown that the first-mentioned formula is sufficient and necessary for the U_t being generated by a suitably chosen group of point-transformations $P \longrightarrow P_t$ (unpublished).

[11] If $||E(\lambda)f||^2$ is differentiable with respect to λ (the case for a type of continuous spectrum defined by Hellinger and Hahn) we have $\int_{-\infty}^{+\infty} \frac{d}{d\lambda} ||E(\lambda) f||^2 e^{it\lambda} d\lambda$. By a well-known theorem on Fourier integrals this has lim equal to zero. Cf. reference 3 above.

$t \longrightarrow \pm\infty$

[12] These PROCEEDINGS, pp. 650–660 (Dec., 1931).

[13] Though it is very probable that the case C or C', in which our theorems hold is the general one for dynamical systems, it is not easy to construct effective examples. (See footnote 10.) An example, recently constructed by v. Neumann, will be published soon; it refers to a two-dimensional flow of the following type: the flow takes place in a rectangle, oriented parallel to the X and Y axes, the upper side of which has been replaced by suitable chosen curve $Y = F(X)$. The flow itself is parallel to the positive Y-axis, and each point X, $F(X)$ has to be identified with the corresponding point $X + \alpha$, 0. (The number $X + \alpha$ is to be taken mod. a, where a is the breadth of the parallelogram in the direction of the X-axis; α is a number incommensurable with a.) If $F(X)$ and α are suitably chosen this flow can be shown to fulfill C (and C').

PHYSICAL APPLICATIONS OF THE ERGODIC HYPOTHESIS

BY J. v. NEUMANN

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY

Communicated January 21, 1932

I. In a recent issue of these PROCEEDINGS[1] the author has obtained a proof of the so-called quasi-ergodic hypothesis. The reader is referred to that paper for the precise formulation of this hypothesis, which plays so important a rôle in the foundations of classical statistical mechanics, and thus in the kinetic theory of gasses; the terminology of that paper will be used throughout this note. The exact statement of the mathematical result obtained in the previous paper by the author is as follows:

Let Ω be either the phase-space Φ of the mechanical system considered, or a sub space of Φ invariant under the transformation ($P \longrightarrow P_t$, P a point of Φ, t the time) induced by the equations of motion.[2] Let dv be the volume element defined in Ω invariant[3] under the transformation $P \longrightarrow P_t$, μN the Lebesgue measure (or weight) of $N(\subset \Omega)$ defined by means of dv : $\mu N = \int_N dv$. Let the time of sojourn of P_τ in N during the time $s < \tau < t$, divided by $t - s$, be denoted by $Z_{s,t}(N; P)$.

Then there exists a function $Z(N; P)$ such that, as $t - s \longrightarrow +\infty$, the function of P $Z_{s,t}(N; P)$ converges, in the sense of "strong convergence" in the space of functions of P, to the limit $Z(N; P)$; that is

$$\lim_{t-s \to +\infty} \int_\Omega | Z_{s,t}(N; P) - Z(N; P) |^2 \, dv = 0. \qquad (1)$$

This property determines $Z(N; P)$, which function, in our previous paper, is studied in more detail and calculated explicitly. The condition for the validity of the so-called quasi-ergotic hypothesis is that $Z(N; P)$ be independent of P; we have shown in our earlier paper that this will be true if and only if there exists in Ω no integral of the equations of motion

other than one almost everywhere constant (for more details, cf. the paper in question).

This result was obtained on the basis of a method due to B. O. Koopman[4] in which the study of dynamical systems is undertaken with the aid of certain unitary and Hermitean functional operators, the spectral resolution of which is made use of systematically. Corresponding with the fact that this method operates with reference to function space, is the fact that our result (1) is in terms of "strong convergence," or convergence in the mean. A natural question to ask is whether our results, e.g., (1), could not be established in terms of convergence almost everywhere in Ω.

Mr. G. D. Birkhoff, to whom we communicated these results orally in October, 1931, has subsequently succeeded in establishing the above surmise by means of an extremely astute method of his own in the domain of point set theory; he has proved, i.e., the existence and equality of the numerical limits

$$\lim_{t \to +\infty} Z_{0,t}(N; P), \qquad \lim_{s \to -\infty} Z_{s,0}(N; P)$$

except on a set of measure zero[5]—which limits, in virtue to (1), will be equal to $Z(N; P)$.

In view of these facts, it is of interest to decide which of the two formulations, (1) or (2), corresponds to the actual physical problem of the ergodic hypothesis. It turns out that the weaker form of statement (1) is sufficient, —that it, indeed, is the precise mathematical equivalent of the physical state of affairs. It is to be noted, further, that the knowledge of the spectral resolution $E(\lambda)$, which is fundamental in Koopman's method,[1,4] enables one to dominate the physical situation here completely; in particular, it furnishes a numerical estimation of the degree of convergence of the limiting process connected with the ergodic hypothesis, whereas Birkhoff's existence proof for (2) is of a non-constructive character.

II. The physical statement of the problem is as follows:

Consider a function $f(P)$ (in Ω) which is a physical quantity referring to the macroscopic state of the mechanical system (e.g., the pressure of a gas, the mean energy per degree of freedom or temperature, etc.). $f(P)$ changes with the time, having at the instant t the value $f(P_t)$, so that if the interval of time $s < \tau < t$ is so short that the time taken in the measurement fills it completely, the quantity measured is not $f(P)$ itself but its time average for $s < \tau < t$, viz., $\frac{1}{t-s}\int_s^t f(P_\tau)\, d\tau$. Is it, then, possible to find a constant C (constant with respect to P, C will naturally depend on f) by which $\frac{1}{t-s}\int_s^t f(P_\tau)\, d\tau$ may be replaced in every application without committing too great an error? (The fact that, if such a C exists

at all, the "micro-canonical" mean $\frac{1}{\mu\Omega}\int_\Omega f(P)\,dv$ may be employed, is evident.)

The criterion for such a possibility is obviously the following: Can a constant C be so determined that the statistical dispersion of $\frac{1}{t-s}\int_s^t f(P_\tau)\,d\tau$ about C is small, i.e.,

$$\int_\Omega \left| \frac{1}{t-s}\int_s^t f(P_\tau)\,d\tau - C \right|^2 dv < \epsilon \qquad (3)$$

for a given $\epsilon > 0$? Now (1) states precisely that this is the case for $t - s$ sufficiently large, in the case where $f(P)$ is taken to be the characteristic function of the set $N \subset \Omega$:

$$f(P) = \chi_N(P)\begin{cases} = 1, \text{ for } P \text{ in } N \\ = 0, \text{ otherwise.} \end{cases}$$

From this the truth of (3) follows for all $f(P)$, (that is:

$$\lim_{t-s\to+\infty} \int_\Omega \left| \frac{1}{t-s}\int_s^t f(P_\tau)\,d\tau - C \right|^2 dv = 0) \qquad (3')$$

(cf. note[1]); for the set of functions $f(P)$ for which it is valid forms a closed linear manifold in function space, and such a manifold contains every $f(P)$ if it contains every $\chi_N(P)$ of finite μN (cf. [1]); the integral of $|f(P)|$ must be finite, as is always the case in the applications.

The statement in terms of probability which corresponds to (2) may be made as follows (where the theorem of Egoroff has been used, in virtue of which any almost everywhere convergent sequence of functions will, for an arbitrary $\epsilon > 0$, converge uniformly except for a set of measure $\leqq \epsilon$[6]):

For every $\epsilon > 0$ and $\delta > 0$ there exists a $T = T(\epsilon, \delta)$ such that the occurrence of $\left| \frac{1}{t}\int_0^t f(P_\tau)\,d\tau - C \right| > \delta$ for any $t >$ is of probability $\leqq \epsilon$.

We have here a refinement of the statement that the statistical dispersion is small; but the latter is quite sufficient for all physical applications.

III. The fact that $t - s$ must be "short" in the physical application and "infinitely long" in the mathematical theorem ($t - s \longrightarrow +\infty$ was premised!) can be explained without inconsistency once the degree of convergence in (3′) is estimated. Such an estimation may be made with the aid of the method of Koopman.

In the notation of the earlier paper, (3′) states that

$$\lim_{t-s\to+\infty} \| \sigma_{t,s} f - E_0 f \|^2 = 0$$

(cf. note,[1] for this and for the following), and the expression $\| \ldots \|^2$ on the left was calculated to be[7]

$$\int_{-\infty}^{-0} + \int_{+0}^{+\infty} \left(\frac{\sin \frac{1}{2}\lambda (t-s)}{\frac{1}{2}\lambda (t-s)} \right)^2 d \| E(\lambda) f \|^2.$$

It is a simple matter to evaluate this expression when $E(\lambda)$ is known. For example, if an interval $-\epsilon < \lambda < \epsilon$ exists, which contains no part of the continuous spectrum, and no point spectrum (besides the simple proper value 0), it is readily seen to be $= \frac{4\|f\|^2}{\epsilon^2 (t-s)^2}$. In the general case various formulae can be obtained, for example,[8]

$$\left\| \left(E\left(\frac{1}{\sqrt{t-s}} \right) - E(+0) \right) f \right\|^2 + \left\| \left(E(-0) - E\left(-\frac{1}{\sqrt{t-s}} \right) \right) f \right\|^2 + \frac{4\|f\|^2}{t-s},$$

which furnishes an immediate evaluation. In the case mentioned first $\frac{\| \sigma_{t,s} f - E_0 f \|}{\| f \|}$ would be $\leqq \frac{2}{\epsilon (t-s)}$ this converging uniformly to 0 for all functions f.[9]

These evaluations could easily be analyzed further, but we shall leave the matter now. The example was only given to illustrate the use of Koopman's method in the setting of physical questions.

[1] *Proc. Nat. Acad. Sci.*, **18**, 1 (1932).

[2] That is, an "integral surface," e.g., an energy surface.

[3] In virtue of Liouville's theorem dv is the usual element of volume in the case $\Omega = \Phi$.

[4] *Proc. Nat. Acad. Sci.*, **17**, 5 (1931).

[5] *Ibid.*, **17**, 12 (1931).

[6] *Paris C. R.*, **152** (1911).

[7] The f appearing there is to be replaced by $f - E_0 f$ (E_0 is permutable with every $E(\lambda)$ and $\sigma_{t,s}$); this has for effect that $\int_{-\infty}^{-0} + \int_{+0}^{+\infty}$ replaces the earlier $\int_{-\infty}^{+\infty}$.

[8] Separate $\int_{-\infty}^{-0} + \int_{+0}^{+\infty}$ into $\left[\int_{+0}^{+\frac{1}{\sqrt{t-s}}} + \int_{-\frac{1}{\sqrt{t-s}}}^{-0} \right] + \left[\int_{+\frac{1}{\sqrt{t-s}}}^{+\infty} + \int_{-\infty}^{-\frac{1}{\sqrt{t-s}}} \right]$ and not that $\left(\frac{\sin \frac{1}{2}\lambda (t-s)}{\frac{1}{2}\lambda (t-s)} \right)^2 \leqq 1$ and $\leqq \left(\frac{2}{\lambda (t-s)} \right)^2$.

[9] It is easily shown, that this only occurs for periodical motions in which all paths have the same period.

COMPLEX GEOMETRY AND RELATIVITY: THEORY OF THE "RAC" CURVATURE

BY EDWARD KASNER

DEPARTMENT OF MATHEMATICS, COLUMBIA UNIVERSITY

Communicated February 3, 1932

1. "If P is a regular point of an analytic curve and if Q approaches P along the curve, the limit of the ratio of the arc PQ to the chord PQ is unity." This familiar assertion is perfectly accurate in the real domain, but when complex points and curves are admitted, it is necessary to make reservations, as the author has indicated.[1] The statement is then true only "in general" and it proves interesting as a curiosity to examine all the possible imaginary exceptions.

I here wish to carry the theory further, especially for curved surfaces, calling attention especially to the fundamental Theorem III, and to emphasize that the exceptions become *real* in the application to *relativity* (four dimensional space of events). Furthermore, certain remarkable discontinuities present themselves, which suggest an interesting analogy to *quantum theory*, rather far-fetched I must admit, and without physical evidence at the present time.

We shall assume our *curves* in the form of power series with integral exponents

$$y = a_1x + a_2x^2 + \ldots,$$

and similarly in space of three or more dimensions, the given point always being taken as origin.

Also our surfaces will be assumed in the (regular analytic) form

$$z = c_{10}x + c_{01}y + c_{20}x^2 + c_{11}xy + c_{02}y^2 + \ldots.$$

The variables, x, y, z and the coefficients a, c, are arbitrary *complex numbers*. It is shown (in section 2) that surfaces fall into five distinct types.

I suggest the name *rac curvature* or more briefly *rac*, for the limiting ratio of arc to chord, that is, in symbols,

$$R = \text{rac} = \text{Limit}\,\frac{\text{arc}}{\text{chord}} = \text{Limit}\,\frac{\alpha}{\gamma}$$

Thus for a given curve at a given point there is a definite R. For ordinary real curves $R = 1$, of course, but for certain types of imaginary curves (namely, those with a minimal tangent line, for example, the imaginary parabola $y = ix + x^2$) R may take other values; the main exceptional value turns out to be

$$\frac{2}{3}\sqrt{2} = 0.9428$$

and this particular irrational number plays a predominating rôle in our theory. Curves can be constructed, however, such that R takes any assigned (real or complex) value, as soon as our space of construction has more than two dimensions.

In general the rac curvature changes discontinuously if either the point moves continuously along the given curve, or if the given curve is varied continuously.

The only case where R does not exist is that where the curve is a minimal straight line, that is, a line whose slope is $\pm i$, for then both arc and chord vanish. We omit this trivial case in the following.

Confining ourselves first to two dimensions, we proved in the former paper that the precise statement is as follows.

THEOREM I. *At any regular point of a plane analytic curve the rac is one of the numbers of the discrete set*

$$1, \frac{2}{3}\sqrt{2}, \frac{2}{4}\sqrt{3}, \frac{2}{5}\sqrt{4}, \frac{2}{6}\sqrt{5}, \ldots, \frac{2}{n+1}\sqrt{n}, \ldots, \qquad (1)$$

whose values to two decimals are

$$1.00,\ 0.94,\ 0.86,\ 0.80,\ 0.74,\ \ldots. \qquad (2)$$

The rac is in fact equal to the n^{th} number of the set at a point where the curve has contact of order $n-1$ with a minimal line.

If the slope is not $\pm i$, that is, if the tangent line is not minimal, the rac will (as in the real case) be unity.

We remark that in the plane all the values of the rac in the countable set (1) are real, positive, and less than one. The set (1) has 0 for its limit, but 0 is not in the set.

In space of more than two dimensions the result is strikingly different.[1]

THEOREM II. *In space of three or more dimensions the rac of a regular analytic curve may have any real or complex value whatever (including 0 and ∞).*

The rac as we have said is usually equal to one. Where the curve follows a minimal direction it is as a rule equal to 0.94, just as inthe plane. The other values are assumed only under special conditions analogous to the requirement in the plane of higher contact with the minimal tangent. The full geometry of the situation is complicated, and will be taken up in another paper. It depends on differential invariants, metric and conformal.

2. In the following we deal with space of three dimensions. We wish to consider the question: "*If P is a regular point of an arbitrary (analytic)*

surface, and if Q is to reach P by moving along a regular path on the surface, what values of the rac can be attained?"

We call the totality of values the *rac set* (of the given surface at the given point).

Of course Theorem I will be part of the answer. However, it must be supplemented, even for planes, in space of three dimensions; for in addition to the *ordinary* (Euclidean or Cartesian) planes contemplated in Theorem I we find in complex space also *minimal* planes. Through each point of an ordinary plane there pass *two* minimal lines of the plane, while a minimal plane has but a *single* minimal line through each point. The distance between two points of a minimal plane is the same, along whatever path in the plane it is measured.[2] The rac therefore is always unity for any curve in such a plane.

For a non-minimal (Euclidean or Cartesian) plane the rac set is

$$R_n = \frac{2\sqrt{n}}{n+1} \tag{3}$$

as described in Theorem I, but for a minimal plane the rac set is merely unity.

Our question about surfaces is answered by the following

FUNDAMENTAL THEOREM III. *Five and only five types of rac set are possible for regular analytic surfaces, namely:*

(*A*) *The totality of complex numbers (including* 0 *and* ∞).

(*B*) *All complex numbers with a single real exception: this exception is one of the odd terms of the set* (1) *other than unity, that is,* R_3, R_5, ...

(*C*) *The discrete set of real numbers* (1)

$$1.00,\ 0.94,\ 0.86,\ 0.80,\ 0.74,\ \ldots.$$

where the general term is

$$R_n = \frac{2\sqrt{n}}{n+1}$$

(*D*) *Just two real numbers: unity and one of the even terms of the set* (1), *that is,* R_2, R_4, ...

(*E*) *Unity alone.*

The possible rac sets have been arranged here according to their magnitude, not according to the generality of their occurrence. The most general or probable type of a random surface is, in fact, type *B*, and, as we shall soon see, the particular species B_3, where the particular omitted value is $R_3 = 0.86$.

THEOREM IV. *At a general point of a general surface the rac can attain any value of the entire complex domain except the particular value* $R_3 =$

$\frac{\sqrt{3}}{2} = 0.86$. *General here means that the tangent plane is non-minimal and the point is non-umbilical (principal radii of curvature unequal).*

We have directly from Theorem III,

THEOREM V. *All possible rac sets, in order of magnitude, can be represented by the five symbols*

$$c,\ c-1,\ \aleph_0,\ 2,\ 1.$$

The appropriateness of the symbols can be seen from the fundamental theorem. Of course, c here denotes the complex continuum. The third type symbol is aleph null.

Some obvious corollaries may be mentioned. If one imaginary rac is possible (at a given point) so is every other. If there are three distinct racs, there are surely an infinite number. If all the racs are real, the set is either finite or denumerable, never continuous.

If we wish to classify surfaces not merely with respect to the bigness of the rac set, but according to the particular values in the rac set, we see that, while types A, C, E are unique or determinate, type B must be subdivided into species which may be denoted by B_3, B_5, ...; and similarly type D must be subdivided into species D_2, D_4, ...

Thus the rac set of B_3 is the entire continuum except $R_2 = 0.86$, and the rac set of D_2 is 1 and $R_2 = 0.94$.

THEOREM VI. *The only possible rac sets of surfaces are those belonging to the species*

$$A;\ B_3,\ B_5,\ \ldots;\ C;\ D_2,\ D_4,\ \ldots;\ E,$$

where B_K *omits the value* $\frac{2\sqrt{K}}{K+1}$, *and* D_K *includes the value* $\frac{2\sqrt{K}}{K+1}$. *Every surface (at a point) belongs to one and only one of these species.*

We add the equations of a few simple algebraic surfaces, displaying at the origin the five types of rac sets.

(A)	$z = x^2 + ixy$, or $z = ix + x^3$	all racs
(B)	$z = x^2$, or $z = ix + xy$	all but 0.86
(C)	$z = x^2 + y^2$, or $z = 0$	the discrete set (I)
(D)	$z = ix + x^2$	1 and 0.94
(E)	$z = ix$	1

In what follows we merely indicate the calculations necessary to establish these theorems, and describe in geometrical terms how the rac changes as the curve through the point of the surface is varied. The tangent plane at the point may of course be either an ordinary plane or a minimal plane; as we shall see, the two cases are essentially different. A full discussion will be published later.

3. Suppose first that the tangent plane at a given point of a surface is

not minimal. Then we may assume it to be the x, y plane, and the equation of the surface with the given point as origin becomes

$$z = c_{20}x^2 + c_{11}xy + c_{02}y^2 + c_{30}x^3 + \ldots . \tag{4}$$

Let p be the lowest exponent of x remaining in (4) when ix is substituted for y, and q the lowest exponent when $y = -ix$. Thus p and q will usually equal 2, but may be any greater integers. If $y = ix$ results in $z = 0$, we shall say that $p = \infty$, and similarly for q.

A curve on the surface regular at the origin can be defined by

$$y = a_1x + a_nx^n + \ldots , \tag{5}$$

$$z = b_2x^2 + b_3x^3 + \ldots . \tag{6}$$

where a_n is written for the first non-zero coefficient after a_1; (as before, n may equal ∞). The b's are determined by substituting (6) into (4).

The chord and the arc are, respectively,

$$\gamma \equiv (x^2 + y^2 + z^2)^{1/2} \text{ and } \alpha \equiv \int_0^x (1 + y'^2 + z'^2)^{1/2}dx. \tag{7}$$

If $1 + a_1^2 \neq 0$, the leading term in the series for γ as well as for α is $(1 + a_1^2)^{1/2}x$. Hence $\alpha/\gamma \longrightarrow 1$. If a_1 equals, say, $+i$, three cases must be made according as n is less than, equal to, or greater than $2p - 1$.

4. We can give a geometrical form to the results.

When the tangent plane at a given point of a surface is not minimal, it contains two distinct minimal lines through the point of contact with the surface. The normal plane through each of these lines is a minimal plane. The two minimal lines of the tangent plane have contact of orders $p - 1$ and $q - 1$, respectively, with the sections of the surface by these normal planes.

If a curve on the surface passes through the given point in a non-minimal direction, the rac as we know must be unity. But consider a curve tangent to the first minimal line. If it diverges too rapidly from the corresponding normal plane, it remains practically in the tangent plane, and therefore assumes one of the early racs of the plane set (I). If it leaves the normal plane too slowly, its initial portion, which determines the rac, is very nearly in the normal plane. Since that plane is minimal, the rac equals unity. But for a certain intermediate order of deviation from the normal plane, the curve is effectively in space. The rac may then have any value at all, with two exceptions: 1 and $\sqrt{2p - 1}/p$. These two are lost, the curves at this stage being neither in the tangent plane nor in the normal plane.

If $p = \infty$, every curve tangent to the minimal line must belong to the first of the three categories described above. Hence these curves contribute just the plane racs (I).

Summing up, we find that near the first minimal line every rac appears except $\sqrt{2p-1}/p$, unless $p = \infty$; then only the set (I) can appear. A corresponding statement holds for the other minimal line; we merely read q for p.

Therefore, if $p \neq q$, every rac will be attained by some curve through the point. If $p = q$, and both are finite, every rac will be possible with the one exception, $\sqrt{2p-1}/p$. If $p = q = \infty$, the racs at the point will be the plane set (I).

It is clear from (4) that in the general case, when the c's are arbitrary, p and q each equal 2. The single missing rac is then $\sqrt{3}/2 = 0.86$, as Theorem IV asserts. The geometric meaning of $p = q = 2$ is that the point is not umbilical (principal radii distinct).

The condition $p = q = \infty$ defines the points at which the two minimal lines of the tangent plane actually lie in the surface. These are always umbilical points. Every point of an ordinary sphere, and of course every point of an ordinary plane, is of this nature.

THEOREM VII. *A point of a surface at which the tangent plane is non-minimal is characterized for our problem by the integers p and q. The rac set must be one of the three types (A), (B) and (C) of Theorem III.*

5. If the tangent plane at a point of a surface is a minimal plane, we may assume that it has the equation $z = ix$. Then the equation of the surface, arranged in powers of y, is

$$z = ix + (c_{k0}x^k + c_{k+10}x^{k+1} + \ldots) + (c_{m1}x^m + \ldots)\,y + \ldots \qquad (8)$$

$$(c_{k0},\, c_{m1} \neq 0;\ 2 \leqq K \leqq \infty,\ 1 \leqq m \leqq \infty.)$$

A curve on the surface is defined by

$$y = a_n x^n + a_{n+1}x^{n+1} + \ldots, \qquad (9)$$

$$z = ix + b_2 x^2 + \ldots . \qquad (10)$$

($a_n \neq 0$; $1 \leqq n \leqq \infty$). Three cases arise, according as k is less than, equal to, or greater than $2m + 1$, and each requires subcases depending on n.

The complicated geometry of the situation will be discussed elsewhere by my student George Comenetz.

THEOREM VIII. *A point of a regular surface at which the tangent plane is minimal is characterized for our problem by the integers k and m. The rac set must be one of the four types A, B, D, E of the fundamental theorem III.*

Theorem III is the sum of Theorems VII and VIII. The examples following Theorem VI have been chosen so as to illustrate every possible type of rac set, for the minimal as well as for the ordinary tangent plane.

6. The limit of the ratio of arc to chord is invariant under conformal

transformations, as might be expected. Our theorems are therefore part of conformal geometry. In fact, the complete conformal group, in complex space, may be defined as consisting of all point transformations (real and complex) for which the rac is invariant. It follows in particular that the rac set at any point of a spherical surface is the same set (1) as for the plane.

7. The application to relativity theory was indicated briefly in a note in *Nature*,[3] published in 1921.

In Minkowski space, the interval between two successive positions of a particle is measured by the formula $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$, where t is the time and c is the velocity of light. Of course, the variables now are real. The interval between two finitely different positions of a particle, or two events, can be measured either along the *arc* or along the *chord;* that is, either along the actual world line or along the world line which the particle would trace if it moved from the first position to the second position with constant velocity, in a three-dimensional straight line. A definite physical meaning is thus given to the limit of the ratio of the arc to the chord, which we have named the *rac*.

8. All our results can be translated for this definition of distance (or interval), with the striking difference that the exceptional cases, which before were all imaginary, now become *real* on account of the minus signs in the interval formula. For example, we can predict that:

If a particle (light photon) starts out with the velocity of light c and then continuously slows down, (the initial acceleration not being zero), the limit of the ratio of the arc to the chord will be $R_2 = \frac{2\sqrt{2}}{3} = 0.94$.

The initial velocity, of course, must equal c with mathematical exactness. If the initial velocity is less than c, then of course $R = 1$. This phenomenon therefore is probably beyond the range of the laboratory, but there is no doubt of its theoretical validity.

The other results of the preceding pages give theorems about rectilinear or planar motion. Theorem I, for instance, states that:

If a particle moves in a straight line in such a way that its displacement is a regular function of the time, its rac at any instant will be one of the numbers 1.00, 0.94, 0.86, 0.80, 0.74, ... *The values other than unity will be attained only at the moments when the particle is moving with the velocity of light.*

The rac will in fact be equal to the nth number of the set R_n whenever the velocity is equal to c for $n - 1$ consecutive instants; that is, whenever the ordinary acceleration and the higher accelerations vanish up to, but not including, the $(n - 1)$th. (It is not necessary to require the motion to be analytic, as we are now in the real domain. The existence and continuity of the derivatives of the displacement suffice.[4])

9. The *discontinuity* which thus presents itself in the rac set of recti-

linear motion suggests at least an analogy with the discontinuous character of energy radiation (quantum theory). Both occur only when the velocity is the velocity of light. If R denotes the rac in (1) we may write

$$\frac{R^2}{4} = \frac{m-1}{m^2} = \frac{1}{m} - \frac{1}{m^2} \tag{11}$$

where $m = n - 1$ is an integer greater than one. Perhaps this formula may in the future be connected with spectrum line series.

10. In the general Einstein relativity theory we deal with a Riemann manifold of four dimensions. The formulas, involving the potentials g_{ik}, are much more complicated, but again we find exceptional values for the rac whenever the initial velocity is that of light—that is, whenever the world line is tangent to a null geodesic on the curved manifold representing the field of gravitation.

The concept of rac can be generalized further so as to apply to other problems in the calculus of variations. Thus for the standard or simplest form of problem,

$$\int F(x, y, y')\, dx = \text{minimum}$$

we take the ratio of the value of the integral along a given curve to the value along the extremal joining the two points, and then take the limit as the second point approaches the first point. The usual value will be unity, but exceptional values can occur when the given curve touches one of the ∞^1 curves $F = 0$. In this problem we assume the ∞^1 curves to be real and work entirely in the real domain, requiring differentiability but not analyticity.

[1] "The Ratio of the Arc to the Chord of an Analytic Curve Need Not Approach Unity," *Bull. Am. Math. Soc.*, **20**, 524–531 (1914).

[2] Coolidge, *Geometry of the Complex Domain*, 1924. In space of four dimensions there exist also *null* planes, where *all* distances vanish, and in higher spaces curved null surfaces and spreads may exist. See my abstracts on null geometry in *Bull. Amer. Math. Soc.*, 1926–27.

[3] "Relativity: Particles Starting with the Velocity of Light." *Nature*, 434 (December, 1921).

[4] This was proved in my seminar by G. Comenetz and A. Epstein. I wish to thank Mr. Comenetz for his valuable assistance in the writing of the present paper and for completing the discussion of surfaces, especially sections 4 and 5. The results were also checked by C. Boyer.

REGULAR FAMILIES OF CURVES. I*

BY HASSLER WHITNEY†

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY

Communicated February 10, 1932

Given a family of curves, such as the paths of particles in a steady motion of a fluid, it is often convenient to introduce a continuous function $x' = g(x, t)$, with the interpretation that the particle at the point x has moved to the point x' after a time t. It is also useful to have cross-sections of the curves; if the curve through any given point has a direction and this direction varies continuously with the point, we can construct a cross-section in the neighborhood of a point by drawing rays at that point perpendicular to the direction of the curve.

In this note we associate with a general point set a function μ; with its help we can, given a family of curves satisfying very weak conditions, define a function and find cross-sections as above. Examples show the usefulness of the function in topology. Conditions under which a family of curves is "regular" will be given later; the question of orienting the curves and of introducing invariant points will also be discussed.

The Function $\mu(S)$.—Let R be any totally bounded[1] metric space (for instance, a bounded part of Euclidean n-space), and let S be any closed[5] subset of R. Immerse any set of $n \geqq 2$ points $p_1, p_2, \ldots, p_n$ in S; with this set of points P_n is associated a number[2] $d(P_n) = \min_{i \neq j} \rho(p_i, p_j)$, the smallest distance between pairs of the points of P_n (which is 0 if some two points coincide). This number $d(P_n)$ is certainly $\leqq \delta(S)$. Hence, for all choices of P_n, there is an upper limit for $d(P_n)$, which we call $\mu_n(S)$:

$$\mu_n(S) = \overline{\lim_{P_n}} d(P_n) = \overline{\lim_{P_n}} \min_{i \neq j} \rho(p_i, p_j). \tag{1}$$

We define $\mu(S)$ by the equation

$$\mu(S) = \sum_{n=2}^{\infty} \frac{\mu_n(S)}{2^{n-1}}. \tag{2}$$

Some obvious properties of the functions are: $\mu_n(S) \geqq \mu_{n+1}(S)$; $\lim_{n \to \infty} \mu_n(S) = 0$; $\mu_2(S) = \delta(S)$; $\delta(S)/2 \leqq \mu(S) \leqq \delta(S)$.

Suppose $S' \subset V_\epsilon(S)$. Take any $n \geqq 2$ and any $P_n' = p_1', \ldots, p_n'$ in S'. Choose p_i in S so that $\rho(p_i, p_i') < \epsilon$, $i = 1, \ldots, n$. Then $\rho(p_i', p_j') < \rho(p_i, p_j) + 2\epsilon$, hence $d(P_n') < d(P_n) + 2\epsilon$, and $\mu_n(S') \leqq \mu_n(S) + 2\epsilon$. Summing over n, we have

$$\mu(S') \leqq \mu(S) + 2\epsilon, \quad \text{if} \quad S' \subset V_\epsilon(S). \tag{3}$$

If S is compact also, we can write $\mu(S') < \mu(S) + 2\epsilon$.

Suppose $S \subset S'$, $S \neq S'$. Let p be a point of S' not in S, at a distance d from S say, and choose n so large that $\mu_n(S) < d$. Take $i > n$ so that $\mu_{i-1}(S) > \mu_i(S)$. Take any set of $i-1$ points P_{i-1} in S. Add the point p, forming a set of i points P_i' in S'. Now $d(P_{i-1}) \leqq \mu_{i-1}(S) < d$; hence $d(P_i') = d(P_{i-1})$, and thus $\mu_i(S') \geqq \mu_{i-1}(S) > \mu_i(S)$. As $\mu_k(S') \geqq \mu_k(S)$ for all k, we have, on summing,

$$\mu(S') > \mu(S) \quad \textit{if} \quad S \subset S',\ S \neq S'. \tag{4}$$

The last two results we can state in the form: $\mu(S)$ *is a continuous and increasing function of* S.

Regular Families of Curves.—We shall use the word *curve* in any of three senses: The topological image of an open or a closed line segment or of a circle. We shall say that a family of curves forming a totally bounded metric space is *regular* if the following condition is satisfied: Given any arc pq of one of the curves C and any $\epsilon > 0$, there is a $\delta > 0$ such that if $p' \subset V_\delta(p)$ lies on a curve C', then there is a point q' on C' such that the arc $p'q'$ has a parameter distance (écart) less than ϵ from the arc pq. By this we mean that there is a homeomorphism between the two arcs such that if r and r' are any two corresponding points, then $\rho(r, r') < \epsilon$. It may be proved easily that in a regular family, no two of the curves intersect.

Take any point p on a curve C of a regular family and consider the arcs pq, q on C variable. For each point q there is a $\mu(pq)$. $\mu(pq)$ increases continuously as q moves away from p; hence for a given number α not too large, there is a point q such that $\mu(pq) = \alpha$. We write then $q = g(p, \alpha)$:

$$q = g(p, \alpha) \textit{ equivalent to } \mu(pq) = \alpha. \tag{5}$$

For given p and α there are either two, one or no corresponding points q—two, one on each side of p, if p is interior to C and α is small enough.

Continuity of g.—Given any point p on C and a point q interior to C, where $\mu(pq) = \alpha$, and given an $\epsilon > 0$, take first $\beta < \epsilon/2$ so small that there are points $q_1 = g(p, \alpha - 3\beta)$ and $q_2 = g(p, \alpha + 3\beta)$, and the arc $q_1qq_2 \subset V_{\epsilon/2}(q)$. Take δ so small that for any point p' in $V_\delta(p)$, there is an arc $p'q_2'$ whose parameter distance from pq_2 is $< \beta$. Using the correspondence thus given between pq_2 and $p'q_2'$, let q_1' be the point corresponding to q_1. Now each point of $p'q_1'$ is within β of the corresponding point of pq_1, and thus $p'q_1' \subset V_\beta(pq_1)$; hence $\mu(p'q_1') \leqq \mu(pq_1) + 2\beta = \alpha - \beta$. Therefore if $q' = g(p', \alpha')$, $|\alpha - \alpha'| < \beta$, the point q' lies beyond q_1'. But also $pq_2 \subset V_\beta(p'q_2')$, and hence $\mu(p'q_2') \geqq \mu(pq_2) - 2\beta = \alpha + \beta$; q' lies therefore before q_2'. It lies consequently on the arc $q_1'q_2'$. But $q_1'q_2' \subset V_\beta(q_1q_2) \subset V_{\beta+\epsilon/2}(q)$; hence $q' \subset V_\epsilon(q)$. We have now proved that *given a point* p, *a direction on the curve* C *through* p, *and an* α *such that* $q = g(p, \alpha)$ *in the given direction from* p *is interior to* C, *then for every* $\epsilon > 0$ *there is a*

$\delta > 0$ such that if $p' \subset V_\delta(p)$ and $|\alpha - \alpha'| < \delta$, then there is a point $q' = g(p', \alpha') \subset V_\epsilon(q)$. If, p and q being fixed, ϵ is taken small enough, then there is such a point q' only in a single direction from p'. The function g is, of course, uniformly continuous over any closed set of points p and closed range of values α, by the Borel theorem.

Cross-Sections.—Given p interior to C, take q and q^* interior to C and on opposite sides of p, and take σ so that $W_{2\sigma}(q)$ and $W_{2\sigma}(q^*)$ do not intersect. Take any ϵ less than any of the numbers 2σ, $\rho[W_{2\sigma}(q), W_{2\sigma}(q^*)]/2$, $\rho[p, W_{2\sigma}(q)]$, $\rho[p, W_{2\sigma}(q^*)]$. If δ_1 is small enough, then any point p' of C' in $W_{\delta_1}(p)$ lies on a unique arc running from $W_{2\sigma}(q)$ to $W_{2\sigma}(q^*)$, this arc being contained within an arc $q'q^{*\prime}$ whose parameter distance from the arc qq^* is $< \epsilon$, with p and p' being corresponding points on the arcs; moreover, this holds only if the arc of C' be taken in the correct orientation. We shall find a $\delta \leqq \delta_1$ and a point set S with the following properties. (1) Each arc through a point p' of $W_\delta(p)$ as above defined contains exactly one point of S, and every point of S lies on such an arc; also $p \subset S$. (2) $S \subset V_\epsilon(p)$. (3) S is closed. The set S we say forms a *cross-section* through p.

We shall now determine a cross-section S. Take $\eta < \epsilon$ so small that if we follow C from p toward q to the first point s in $W_{2\eta}(q)$, then $sq \subset V_\sigma(q)$. Given any point $p^{(i)}$ near p on the curve $C^{(i)}$, let $r^{(i)}$ be the first point of $C^{(i)}$ starting from $p^{(i)}$ in $W_\eta(q)$; in particular, r is that point on C. Put $U = W_{2\sigma}(q)$ and say $\mu(U \dot{+} pr) = \alpha$. Take $\beta < \eta/2$ so small that if p_1 and p_2 are the points on C such that $\mu(U \dot{+} p_1r) = \alpha - 2\beta$, $\mu(U \dot{+} p_2r) = \alpha + 2\beta$, then $p_1pp_2 \subset V_{\epsilon/2}(p)$.

Take γ_1 so small that if $p_1' \subset V_{\gamma_1}(p_1)$, then $\rho[g(p_1, t), g(p_1', t)] < \beta$, $0 \leqq t \leqq \mu(p_1q)$. We shall show that $U \dot{+} p_1'r' \subset V_\beta(U \dot{+} p_1r)$; it will follow that

$$\mu(U \dot{+} p_1'r') \leqq \mu(U \dot{+} p_1r) + 2\beta = \alpha. \tag{6}$$

It is sufficient to show that $p_1'r' \subset V_\beta(U \dot{+} p_1r)$. If x' is any point of $p_1'r'$ and $x' = g(p_1', t)$, put $x = g(p_1, t)$, using the same t. We shall show that $x \subset U \dot{+} p_1r$; then $\rho(x, x') < \beta$, and the inclusion follows. If $q_1' = g(p_1', t_1)$, where $t_1 = \mu(p_1q)$, then $\rho(q, q_1') < \beta < \eta$. Thus q_1' lies in $V_\eta(q)$, hence r' and thus x' lies between p_1' and q_1', and hence x exists and lies in p_1q; but $p_1r \subset U \dot{+} p_1r$, and $rq \subset sq \subset U$.

A similar argument shows that for some γ_2 corresponding to the same β, if $p_2' \subset V_{\gamma_2}(p_2)$, then

$$\mu(U \dot{+} p_2'r') \geqq \alpha. \tag{7}$$

This will follow when we have shown that $U \dot{+} p_2r \subset V_\beta(U \dot{+} p_2'r')$, or, $p_2r \subset V_\beta(U \dot{+} p_2'r')$. Take $x \subset p_2r$, say $x = g(p_2, t)$, and put $x' = g(p'_2, t)$; then $\rho(x, x') < \beta$. If x' is in $p_2'r'$, $x \subset V_\beta(p_2'r')$; if it is not, it is on C' beyond r'. Then if $s_2 = g(p_2, t_2)$, where $t_2 = \mu(p_2'r')$, x is beyond

s_2. Now $\rho(s_2, r') < \beta < \eta$, and thus $s_2 \subset V_{2\eta}(q)$; hence, by the choice of η, $s_2 r \subset s_2 q \subset V_\sigma(q)$, and thus $x \subset U$, and $x' \subset V_\beta(U)$.

If now we take a $\delta \leqq \delta_1$ so small that any point in $W_\delta(p)$ lies on an arc A joining $W_{2\sigma}(q)$ and $W_{2\sigma}(q^*)$ which cuts both $V_{\gamma_1}(p_1)$ and $V_{\gamma_2}(p_2)$, then on account of (6) and (7), there will be a point p' of A for which $\mu(U + p'r') = \alpha$. The set S of all these points p' forms a cross-section through p, as we now show.

Each arc passing through $W_\delta(p)$ contains just one point of S. Each point p' of S lies on an arc $p_1'p_2'$ in an arc A passing through $W_\delta(p)$; as the parameter distance of $p_1'p_2'$ from p_1p_2 is $< \beta < \epsilon/2$ and $p_1p_2 \subset V_{\epsilon/2}(p)$, $p' \subset V_\epsilon(p)$. We must show still that S is closed; this will follow if we show that for any point p' of C' in S, the points of S on all arcs passing through some $V_{\delta'}(p')$ lie in a $V_{\epsilon'}(p')$. If p' is p, this follows directly from the proof we have given. If not, we replace in the proof C by C', C' by a C'', p_1 and p_2 by p_3 and p_4 close to p', etc. We can use the same β as before, when we notice one fact: there is a point q_2' of C' in $V_\beta(q)$ such that $r'q_2' \subset U$. For put $q_2' = g(p_2', t)$, where $t = \mu(p_2q)$; then $q_2' \subset V_\beta(q)$. We saw that $s_2 \subset V_{2\eta}(q)$ and thus $s_2q \subset V_\sigma(q)$; but $r'q_2' \subset V_\beta(s_2q) \subset V_{2\sigma}(q)$. The rest of the proof follows the argument already given, and S is shown to be a cross-section.

Examples.—(*a*) A theorem of R. L. Moore[3] can be proved simply with the aid of the function μ. It is not hard to show that his family of arcs is a regular family. Map the arc AB onto the segment (0, 0), (1, 0) of the x,y-plane, and map each point p of an arc A_1D_1 of the family onto the point (x, y), where $y = \mu(A_1p)/\mu(A_1D_1)$, if A_1 was mapped onto $(x, 0)$. Evidently $\mu(A_1D_1)$ is a continuous function of the point A_1; the rest of the argument follows from the continuity of the function g. The theorem holds if the arc AB is replaced by a more general point set.

(*b*) We can also prove a theorem of Kerekjarto.[4] A cross-section of the given closed curves is in this case seen to be an arc; a denumerable number of such arcs may be put together to form an arc from the outer curve to the innermost point. Map this arc onto the radius of a circle, and map the closed curves onto concentric circles as in the example above.

* Presented to the American Mathematical Society, March 25, 1932.

† NATIONAL RESEARCH FELLOW.

[1] Total beschränkt (Hansdorff). This is equivalent to asserting that

$$\lim_{n \to \infty} \mu_n(s) = 0.$$

[2] $\rho(p, q)$ is the distance between two points p and q, or in general, between two point sets; $\delta(S)$ is the diameter of S; $V_\epsilon(S)$ consists of those points q for which $\rho(q, S) < \epsilon$, and $W_\epsilon(S)$, those points q for which $\rho(p, S) \leqq \epsilon$.

[3] R. L. Moore, *Fund. Math.*, **4**, 106–117(1923). Theorem I.

[4] v. Kerekjarto, *Topologie* I, p. 246. See also pp. 238–239.

[5] I. E., any limit points of a sequence in S there may be in R lie also in S.

RECENT CONTRIBUTIONS TO THE ERGODIC THEORY

BY G. D. BIRKHOFF AND B. O. KOOPMAN

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY

Communicated February 13, 1932

I. *The Original Formulations.*—Early in the kinetic theory of matter it was recognized that the replacement of mean values in time by mean values in phase-space is an indispensable operation. In order to justify this substitution, recourse was had by Boltzmann and Maxwell to their celebrated Ergodic Hypothesis, which states that, in the systems considered in the kinetic theory, each particular motion will pass, sooner or later, through every state of motion, or phase (combination of position and velocity), consistent with its energy; and (implicitly) that the length of time during which it exhibits a given set of phases is proportional to the relative abundance of the latter. In other words, each path-curve in phase-space passes through every point of its energy hypersurface, remaining in a given region of the latter for a length of time proportional to the extent (hypervolume) of the region. It was recognized by Boltzmann and Maxwell that account must be taken of all possible single-valued integrals of ordinary type, thus reducing the dimensionality of the phase-space.

The extreme unlikelihood of the Ergodic Hypothesis, owing to the presence of special periodic motions, was pointed out by Kelvin.[1] Its mathematical impossibility was apparent to Poincaré,[2] who indicated the only possible direction of effective modification, namely, the later Quasi-Ergodic Hypothesis of P. and T. Ehrenfest:[3]

Let Ω be the hypersurface in phase-space corresponding to the given value of the energy of the system, and suppose its extent to be finite. Then each non-exceptional path-curve in Ω passes through every region of Ω of positive volume, remaining there for an average time equal to the ratio of this volume to that of Ω.

Furthermore, P. and T. Ehrenfest[4] observed that even if the system is quasi-ergodic in the sense that everywhere dense path-curves exist, the mean time of sojourn along the path-curve through P in a given region M may vary discontinuously from path to path.

For the detailed history of the qualifications and speculations, the reader is referred to the "Encyklopädie der Mathematischen Wissenschaften."[5]

II. *Introduction of the Modern Theory of Real Variables into Dynamics.*—In his discussion of recurrent motion, H. Poincaré[6] introduces the fundamental notion of a dynamical property which, without being true for all possible motions, has a probability of one of being realized. Poincaré wrote before Lebesgue's great work, but the very steps of his proof, as well as the formulation of his theorem, are all in almost an exact form for

interpretation in terms of the theory of measure. Such an interpretation was accomplished by C. Carathéodory,[7] who renders "exception of probability zero" as "exceptions forming in Ω a set of measure zero;" and such is the first entrance into the realm of dynamics of the modern theory of real variables.

Shortly afterward G. D. Birkhoff[8] conjectured that in systems having everywhere dense path curves the exceptional path curves (i.e., those not everywhere dense in Ω) are of Lebesgue measure zero—a conjecture made also by A. Smekal[9] in connection with the Quasi-Ergodic Hypothesis.

A few years later, the notion of *metrical transitivity* was introduced in another connection by G. D. Birkhoff and P. A. Smith.[10] If $(P \longrightarrow P_t)$ is a continuous one-parameter group of automorphisms of the metrical space Ω, it is said to be metrically transitive if and only if Ω cannot be decomposed into two subsets each of measure greater than zero, and each invariant under every transformation of the group.

The fundamental rôle to be played by this concept in the ergodic theory will become evident by the developments described below.

A further application of modern analysis to dynamics has recently been made by B. O. Koopman;[11] the transformation-group $(P \longrightarrow P_t)$ in Ω is interpreted as a linear functional operator (the U_t-operator: $U_t f(P) = f(P_t)$) which, when regarded in the space $\mathfrak{L}_2$ of Lebesgue-measurable functions of summable square, becomes unitary in the case where $(P \longrightarrow P_t)$ has a positive integral invariant. This U_t-operator is then studied by means of its spectral resolution $E(\lambda)$. In this study the notion of non-integrability in $\mathfrak{L}_2$ is used, which is equivalent to metrical transitivity when the measure of Ω is finite.[12]

All the results described in this section concern properties true "almost everywhere" in the sense of Lebesgue measure.

III. *The Mean Ergodic Theorem.*—The first one actually to establish a general theorem bearing fundamentally on the Quasi-Ergodic Hypothesis was J. v. Neumann,[13] who, with the aid of the above theory of the U_t-operator, proved what we will call the Mean Ergodic Theorem, to the following effect:

Under the above hypotheses, if $Z_{\alpha,\beta}(P;\ M)$ denotes the relative sojourn of the moving point $P_t (P_0 = P)$ in M in the time interval $\alpha \leqq t \leqq \beta$ (i.e., it is the length of this sojourn divided by $\beta - \alpha$), then, as $\beta - \alpha \longrightarrow +\infty$, M remaining fixed, $Z_{\alpha,\beta}(P;\ M)$ *converges in the mean to a limit* $Z(P;\ M)$ (a limiting process involving neighboring motions). Furthermore, $Z(P;\ M)$ is independent of P if and only if the system is non-integrable in $\mathfrak{L}_2$,[12] in which case $Z(P;\ M)$ = measure of M/measure of Ω, when Ω is of finite measure, and = zero otherwise. When the system is integrable in $\mathfrak{L}_2$, $Z(P;\ M)$ is given a simple expression in terms of generalized derivatives.

This theorem is important not merely as the first general mathematically rigorous treatment of the question, but because it is sufficient for the needs of the kinetic theory (if metrical transitivity is granted): From the standpoint of statistical mechanics, what is needed, as v. Neumann has observed,[14] is the knowledge that, statistically speaking, time-means of function on Ω can be replaced by space-means in Ω; and the convergence in the mean established by v. Neumann is the precise statement that the dispersion or "standard deviation" of the former from the latter vanishes when the time-interval is sufficiently large. Furthermore, v. Neumann's method of proof provides a means of computing a lower limit for the interval of time necessary to make the dispersion less than a given $\epsilon > 0$.

As subsequent developments along these lines, we mention the simplified treatment given by E. Hopf,[15] who makes no use of the spectral resolution $E(\lambda)$, only employing the simpler properties of the unitary operator U_t.

IV. *The Ergodic Theorem.*—On October 22, 1931, Mr. v. Neumann communicated personally to the authors of this note his results on the Mean Ergodic Theorem. As Mr. v. Neumann pointed out then, this positive theorem raised at once the important question as to whether or not ordinary time means exist along the individual path-curves (point convergence of $Z_{\alpha,\beta}(P;\ M)$) excepting for a possible set of Lebesgue measure zero. Smekal[16] had inclined to the opinion that this is probably not the case.

Shortly thereafter, G. D. Birkhoff[17] obtained by entirely new methods the following result, which we will call the Ergodic Theorem:

Under the same conditions as before, $Z_{\alpha,\beta}(P;\ M) \longrightarrow Z(P;\ M)$ as $\beta - \alpha \longrightarrow +\infty$, for all P of Ω except for a set of measure zero.

With regard to the scope of this theorem, we may make the following remarks:

1. From the point of view of the gross statistics on Ω (classical kinetic theory), it is equivalent in its implications to the Mean Ergodic Theorem.

2. From the viewpoint of the detailed statistics along an individual path-curve, it is fundamentally more far-reaching: in it is proved for the first time that the relative time of sojourn along almost every individual path-curve *exists*, a result often assumed implicitly in the writing of physicists, but never proved.

3. Whereas the Mean Ergodic Theorem belongs in its very nature to the theory of unitary U_t-operators in $\mathfrak{L}_2$, and had always been proved by this technique, the Ergodic Theorem steps outside this domain, and was proved by Birkhoff with the sole use of his fundamental lemma, in which the barest notion of the U_t-operator is employed:

Lemma.—Let T be a measure-preserving automorphism of the metrical space Ω, and let $f(P)$ be any measurable function defined on Ω. Then, as $n \longrightarrow \pm\infty$.

$$\frac{f(P) + f(TP) + \ldots + f(T^{n-1}P)}{n}$$

converges almost everywhere on Ω to a limit $f_1(P)$.

It may be stated in conclusion that the outstanding unsolved problem in the ergodic theory is the question of the truth or falsity of metrical transitivity for general Hamiltonian systems. In other words, the *Quasi-Ergodic Hypothesis* has been replaced by its modern version: the *Hypothesis of Metrical Transitivity*.

[1] *Collected Works*, Vol. 4, 484–512. This was published first in 1891.

[2] *Revue Générale des Sciences*, 516 (1894), et. seq. The impossibility of the Ergodic Hypothesis was proved by Plancherel and Rosenthal, *Ann. phys.*, [4] **42**, 796, 1061 (1913). The Ergodic Hypothesis is called by Maxwell and his English-speaking followers the "Principle of the Continuity of Path."

[3] *Encykl. d. Math. Wissenschaften*, **4**, Art. 32; Anm. 89a and 90 (1911). The possibility of exceptional path curves is not taken account of here.

[4] *Ibid.*, Anm. 93.

[5] *Ibid.*, **4**, Art. 32; Nr. 10a (1911), and **5**, Art. 28 (1923).

[6] *Les Méthodes Nouvelles de la Mécanique Celeste*, Paris, Gauthier-Villars (1899), Vol. 3, Chap. XXVI: Stabilité á la Poisson.

[7] *Berlin, Ber.*, 580 (1919).

[8] *Acta Math.*, **43**, 113 (1922).

[9] *Encykl. d. Math. Wissenschaften*, **5**, Art. 28, p. 869, Anm. 8 (1926).

[10] "Structure Analysis of Surface Transformations," *J. Math.*, **7**, 365 (1928). In certain earlier papers this is referred to under the name of "strong transitivity" (cf. infra, 17, 19).

[11] "Hamiltonian Systems and Hilbert Space," these PROCEEDINGS, **17**, 315–318 (1931).

[12] The system is said to be integrable in $\mathfrak{L}_2$ if and only if a function $f(P)$ on Ω exists which is not almost everywhere constant and such that, for all t, and almost all P of Ω, $f(P_t) = f(P)$. When this is the case, the subsets of Ω for which $f(P) \; (\leqq) \; \lambda$ and for which $f(P) \; (\geqq) \; \lambda$ will each be, for some λ, subsets of Ω of positive measure which remain invariant under $(P \longrightarrow P_t)$ for all t except for sets of measure zero (dependent on t). Furthermore, since $f(P)$ is in $\mathfrak{L}_2$, the measure of the set where $f(P) \; (\geqq) \; \lambda > 0$ (etc.), is finite. Thus, when Ω is of infinite measure, non-integrability in $\mathfrak{L}_2$ and metrical transitivity are not co-extensive, since the latter could occur by the decomposition of Ω into two invariant subsets each of infinite measure.

[13] "Proof of the Quasi-Ergodic Hypothesis," these PROCEEDINGS, **18**, 70–82 (1932).

[14] "Physical Applications of the Quasi-Ergodic Hypothesis," these PROCEEDINGS, (March, 1932).

[15] "On the Time Average Theorem in Dynamics," these PROCEEDINGS, **18**, 93–100 (1932).

[16] *Handbuch der Physik* (H. Geiger and K. Scheel), Vol. 9, 1926, in particular p. 182. It may be observed here that all requisite conditions (existence of time means, etc.), have been known to be fulfilled by conditionally periodic and other similar systems. In this connection see A. Wintner, these PROCEEDINGS, **18**, 248–251 (March, 1932).

[17] "Proof of a Recurrence Theorem for Strongly Transitive Systems, and Proof of the Ergodic Theorem," these PROCEEDINGS, **17**, 650–660 (1931).

ON THE NUMERATORS OF THE CONTINUED FRACTION

$$\frac{\lambda_1}{|x - c_1|} - \frac{\lambda_2}{|x - c_2|} - \cdots\cdots$$

BY J. SHOHAT AND J. SHERMAN

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF PENNSYLVANIA

Communicated January 21, 1932

1. It is known that the infinite continued fraction

$$K \equiv \frac{\lambda_1}{|x - c_1|} - \frac{\lambda_2}{|x - c_2|} - \cdots \quad (\lambda_i, c_i \text{ real};\ \lambda_i > 0) \tag{1}$$

is "associated" with one and only one "positive definite" power series[1,2]

$$P\left(\frac{1}{x}\right) \equiv \sum_{i=0}^{\infty} \frac{\alpha_i}{x^{i+1}} \text{ with } \| \alpha_{i+j} \|_{i,j=0}^{n} > 0 \ (n \geqslant 1) \tag{2}$$

and with at least one Stieltjes integral of the form

$$F(x) \equiv \int_{-\infty}^{\infty} \frac{d\psi(\gamma)}{x-\gamma}, \tag{3}$$

where $\psi(\gamma)$ is bounded and non-decreasing in $(-\infty, \infty)$, with infinitely many points of increase [$\psi(-\infty) = 0$!] and is a solution of the "$[\alpha_n]_0^\infty$ − moment problem," i.e.,

$$\int_{-\infty}^{\infty} x^n d\psi(x) = \alpha_n \ (n^{th} \text{ moment}) \ (n = 0, 1, 2, \ldots).^* \tag{4}$$

The interval $(-\infty, \infty)$ in (4) may be "reducible," i.e., it may reduce to a sub-interval (a, b), if $\psi(x) = \psi(a)$ for $x < a$ and $\psi(x) = \psi(b)$ for $x > b$. Denote by $\frac{\Omega_n(x)}{\phi_n(x)}$ $(n = 0, 1, 2, \ldots)$ the successive convergents to (1). From the very definition of "association" of continued fractions of type (1) with power series (2) and integrals (3), namely,

$$\sum_{i=0}^{\infty} \frac{\alpha_i}{x^{i+1}} - \frac{\Omega_n(x)}{\phi_n(x)} = \frac{\alpha'}{x^{2n+1}} + \frac{\beta'}{x^{2n+2}} + \cdots, \tag{5}$$

we derive the *fundamental orthogonality relations characterizing the* $\phi_n(x)$:

$$\int_{-\infty}^{\infty} \phi_m(x)\ \phi_n(x)\ d\psi(x) = 0 \quad (m \neq n;\ m, n = 0, 1, 2, \ldots). \tag{6}$$

The object of this Note is to state some general properties of the numera-

* We may have, in particular, $d\psi(x) = P(x)dx$, $P(x) \geq 0$.

tors $\Omega_n(x)$, which so far have been studied in very special cases only. The most remarkable property is expressed in

THEOREM I. *The numerators* $\Omega_n(x)$, $(n = 1, 2, \ldots)$ *of the successive convergents to the continued fraction* (1) *form, like the denominators, an orthogonal system of polynomials, i.e., there exists a function* $\psi_1(x)$ *of the same nature as* $\psi(x)$ *in* (4), *such that*

$$\int_{-\infty}^{\infty} \Omega_m(x)\,\Omega_n(x)\,d\psi_1(x) = 0 \quad (m \neq n;\ m, n = 0, 1, 2, \ldots),$$

the interval $(-\infty, \infty)$ *perhaps being "reducible" for* $\psi_1(x)$.

In fact, we have the following easily provable

LEMMA. *Denote by* $\frac{A_n'}{B_n'}$ *and* $\frac{A_n''}{B_n''}$, *respectively, the successive convergents to*

$$K' \equiv \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots, \quad K'' \equiv \frac{a_2|}{|b_2} + \frac{a_3|}{|b_3} + \cdots \quad (a_i, b_i \text{ arbitrary}, a_i \neq 0).$$

Then $A_{n+1}'' = a_1 B_n'$ $(n = 0, 1, 2, \ldots)$. It follows that the polynomials $\Omega_n(x)$ are, disregarding constant factors, the denominators of the successive convergents to

$$K_1 \equiv \frac{\lambda_2\ |}{|x - c_2} - \frac{\lambda_3\ |}{|x - c_3} - \cdots, \tag{7}$$

which, being of the same type as (1), gives rise to relations similar to (5, 6), and this proves our statement.

We may write, without loss of generality,

$$F(x) \equiv \int_{-\infty}^{\infty} \frac{d\psi(\gamma)}{x - y} \sim \sum_{i=0}^{\infty} \frac{\alpha_i}{x^{i+1}} \sim \frac{1\ |}{|x - c_1} - \frac{\lambda_2\ |}{|x - c_2} - \cdots$$

$$(\alpha_i = \int_{-\infty}^{\infty} x^i\,d\psi(x);\ \alpha_0 = 1) \tag{8}$$

$$F_1(x) \equiv \int_{-\infty}^{\infty} \frac{d\psi_1(\gamma)}{x - y} \sim \sum_{i=0}^{\infty} \frac{\beta_i}{x^{i+1}} \sim \frac{1\ |}{|x - c_2} - \frac{\lambda_3\ |}{|x - c_3} - \cdots$$

$$(\beta_i = \int_{-\infty}^{\infty} x^i\,d\psi_1(x);\ \beta_0 = 1), \tag{9}$$

and we thus get (as in (6)):

$$\int_{-\infty}^{\infty} \sigma_m(x)\,\sigma_n(x)\,d\psi_1(x) = 0 \quad (m \neq n;\ m, n = 0, 1, 2, \ldots);$$

$$\sigma_n(x) \equiv \Omega_{n+1}(x) = x^n + \ldots;\ n > 0 \tag{10}$$

We also get, introducing the "normalizing factors"[3]

$$a_n = (1 \cdot \lambda_2 \cdot \lambda_3 \ldots \lambda_{n+1})^{-1/2};\ a_n' = (1 \cdot \lambda_3 \cdot \lambda_4 \cdots \lambda_{n+2})^{-1/2} = \sqrt{\lambda_2}\,a_{n+1}: \tag{11}$$

$$\int_{-\infty}^{\infty} \varphi_m(x)\,\varphi_n(x)\,d\psi(x) = \int_{-\infty}^{\infty} X_m(x)\,X_n(x)\,d\psi_1(x) = \delta_{m,n}\ (m, n = 0, 1, 2, \ldots)$$

$$\varphi_n(x) \equiv a_n \phi_n(x), \quad X_n(x) \equiv a_n' \sigma_n(x). \quad (12)$$

2. Hereafter $\psi(x)$ denotes generally a solution of the $[\alpha_n]_0^\infty$ moment problem related, in view of (5), to K in (1). We need some properties of the zeros of $\phi_n(x)$ and $\Omega_n(x)$ which we denote, respectively, by x_i, x_j' $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, n-1)$.[1,2,4,5]

(i) $x_1 < x_1' < x_2 < x_2' < \ldots < x_{n-1} < x_{n-1}' < x_n$

(ii) As n increases, x_n increases and x_1 decreases. Hence

$$\lim x_1 = \lambda(\geqslant -\infty), \lim x_n = L(\leq +\infty)\ (n \longrightarrow \infty) \quad (13)$$

implies the existence of a solution $\psi(x)$ of the $[\alpha_n]_0^\infty$ moment problem, which is constant in each of the intervals $(-\infty, \lambda)$, (L, ∞)

In other words, (13) implies

$$\int_{\lambda}^{L} \phi_m(x)\,\phi_n(x)\,d\psi(x) = 0 \quad (m \neq n;\ m, n = 0, 1, 2, \ldots) \quad (14)$$

On the other hand, a solution of the same moments problem, say, $\Psi(x)$ which is constant in $(-\infty, \lambda')$ and (L', ∞), with $(\lambda', L') \subset (\lambda, L)$, is impossible, for otherwise, *all* zeros of *any* $\phi_n(x)$ $(n = 1, 2, \ldots)$ lie in (λ', L'), contradicting (13). Thus we are justified in *calling the interval*, (λ, L), *uniquely determined by the limiting values, for* $n \longrightarrow \infty$, *of the extreme zeros of* $\phi_n(x)$, *the "true" interval of orthogonality of the set* $\{\phi_n(x)\}$.

(iii) If (λ, L) is finite, then, in general, infinitely many of the zeros x_i have for limits λ and L, respectively.*

Furthermore, for n sufficiently large, the x_i are everywhere dense in any sub-interval (α, β), which does not contain an interval of constancy of $\psi(x)$.

(iv) If (λ, L) is infinite, the distribution of the x_i depends upon the "determined" or "indetermined" character of the $[\alpha_n]_0^\infty$ -moment problem, i.e., whether it has one or several solutions. If $(\lambda, L) = (\lambda, \infty)$, ($\lambda$ finite) or $(-\infty, L)$ (L finite), or $(-\infty, \infty)$, correspondingly infinitely many x_i have for limit $+\infty$, or $-\infty$, or both.

(v) Whether (λ, L) is finite or infinite, if α and β are points of increase for $\psi(x)$, which is constant in $(\alpha + 0, \beta - 0)$, then $\phi_n(x)$ has in (α, β) at most one zero.

On the basis of these properties we state:

THEOREM II. *Let* (λ, L) *and* (λ', L') *denote the "true" intervals of ortho-*

* In some cases, $\psi(x)$ being a step-function, we may have,[5] as $n \longrightarrow \infty$ $x_n \longrightarrow L$, $x_{n-1} \longrightarrow L' < L$, and so on.

gonality for $\{\phi_n(x)\}$ *and* $\{\Omega_n(x)\}$, *respectively.* (*i*) $(\lambda', L') \subset (\lambda, L)$. (*ii*) *If* (λ, L) *is finite, then, in general,* $\lambda' = \lambda, L' = L$. *Any sub-interval of* (λ, L), *which does not contain an interval of constancy of* $\psi(x)$ *possesses the same property with regard to* $\psi_1(x)$. (*iii*) *If* (λ, L) *is infinite, so is* (λ', L'); *more precisely, if* $(\lambda, L) = (\lambda, \infty)$ *or* $(-\infty, L)$, (λ *or* L *finite*), *then, respectively,* $L' = \infty$ *or* $\lambda' = -\infty$; *if* $(\lambda, L) \equiv (-\infty, \infty)$, *then* $(\lambda', L') \equiv (-\infty, \infty)$.

Illustration: Take $(a, b) \equiv (-1, 1)$, $d\psi(x) = \dfrac{dx}{\sqrt{1-x^2}}$. Then, disregarding constant factors and letting $x = \cos\varphi$,[6]

$$\phi_n(x) = \cos n\varphi; \quad \Omega_n(x) = \sin n\varphi/\sin\varphi$$

$$\int_{-\infty}^{\infty} \sqrt{1-x^2}\,\Omega_n(x)\,\Omega_m(x)\,dx = 0 \qquad (m \neq n;\ m, n = 0, 1, 2, \ldots).$$

3. We conclude with:

THEOREM III. *The* $[\alpha_n]_0^\infty$ *-moment problem related to* K *in* (1) *and the* $[\beta_n]_0^\infty$ *-moment problem related to* K_1 *in* (7) *are either both determined or both indetermined.*

The proof is based upon the following properties of K in (1) and $F(x)$ in (3).[1,2]

(i) For any real t the zeros of the denominators of the continued fraction $K_n(x, t) = \dfrac{\lambda_1}{|x - c_1|} - \dfrac{\lambda_2}{|x - c_2|} - \cdots - \dfrac{\lambda_{n-1}}{|x - c_{n-1}|} - \dfrac{\lambda_n}{|x - c_n + t}$ (λ_i, c_i real; $\lambda_i > 0$) are real and distinct.

(ii) $F(x) = \displaystyle\int_{-\infty}^{\infty} \frac{d\psi(\gamma)}{x-\gamma}$, ($\psi(x)$ as described above) and $\dfrac{1}{F(x)}$ are both analytic in any finite closed domain B of the complex x-plane, which does not contain any point of the real axis.

(iii) From any infinite sequence of convergents to the continued fraction (1) we can extract a sub-sequence which converges to a Stieltjes integral of the form (3) for all non-real x.

(iv) The $[\alpha_n]_0^\infty$ -moment problem related to K in (1) is determined if and only if (1) is completely convergent, i.e., given a domain B as described above, and an arbitrary $\epsilon > 0$, we can find a positive integer N, depending on ϵ and B only, so that $|F(x) - K_n(x, t)| < \epsilon$ ($n \geqslant N$), (t real, arbitrary; x in B). These properties, combined with the relation

$$F(x) = \frac{\lambda_1}{x - c_1 - F_1(x)}, \tag{15}$$

where $F(x)$ and $F_1(x)$ stand, respectively, for the integrals of the form (3)

associated with the continued fractions (1) and (7)*, yield an easy proof of Theorem III.

Further properties of the $\Omega_n(x)$, in particular when the $\phi_n(x)$ are the classical orthogonal polynomials (of Jacobi, Laguerre, Hermite), the explicit expression of $\psi_1(x)$, application of $\Omega_n(x)$ to mechanical quadratures, etc., will be developed by Mr. Sherman elsewhere.

* Due to convergence, we have here not mere "associations" but actual equality.

[1] O. Perron, *Die Lehre von den Kettenbrüchen*, 2d ed. (1929), Ch. VIII, IX.

[2] H. Hamburger, "Ueber eine Erweiterung des Stieltjesschen Momentenproblem, I, II, III," *Math. Annal.*, **81–82**, 235–319, 120–164, 168–187 (1920–21).

[3] Jacques Chokhate (J. Shohat), "Sur le développement de l'intégrale $\int_b^a \frac{p(y)dy}{x-y}\ldots$," *Rend. Circolo Matematico Palermo*, **47**, pp. 25–47 (1923).

[4] Stieltjes, "Quelques recherches sur les quadratures dites mécaniques," *Oeuvres*, **1**, 377–396.

[5] Idem, "Recerches sur les fractions continues," *Ibid.*, **2**, pp. 402–566.

[6] C. Possé, "Sur quelques applications des fractions continues algébriques," (St. Petersbourg, 1886), Ch. IV.

ERRATA

On page 132 of the January issue there was omitted one full line which should have appeared as the second line of that page, making the first and second sentences of the page read:

The technique was varied without significant differences in the results. When the eggs in the alkaline sea water were returned to normal sea water after varying intervals the only change noted was a greater number of normal larvae and not a larger total number of developing eggs.